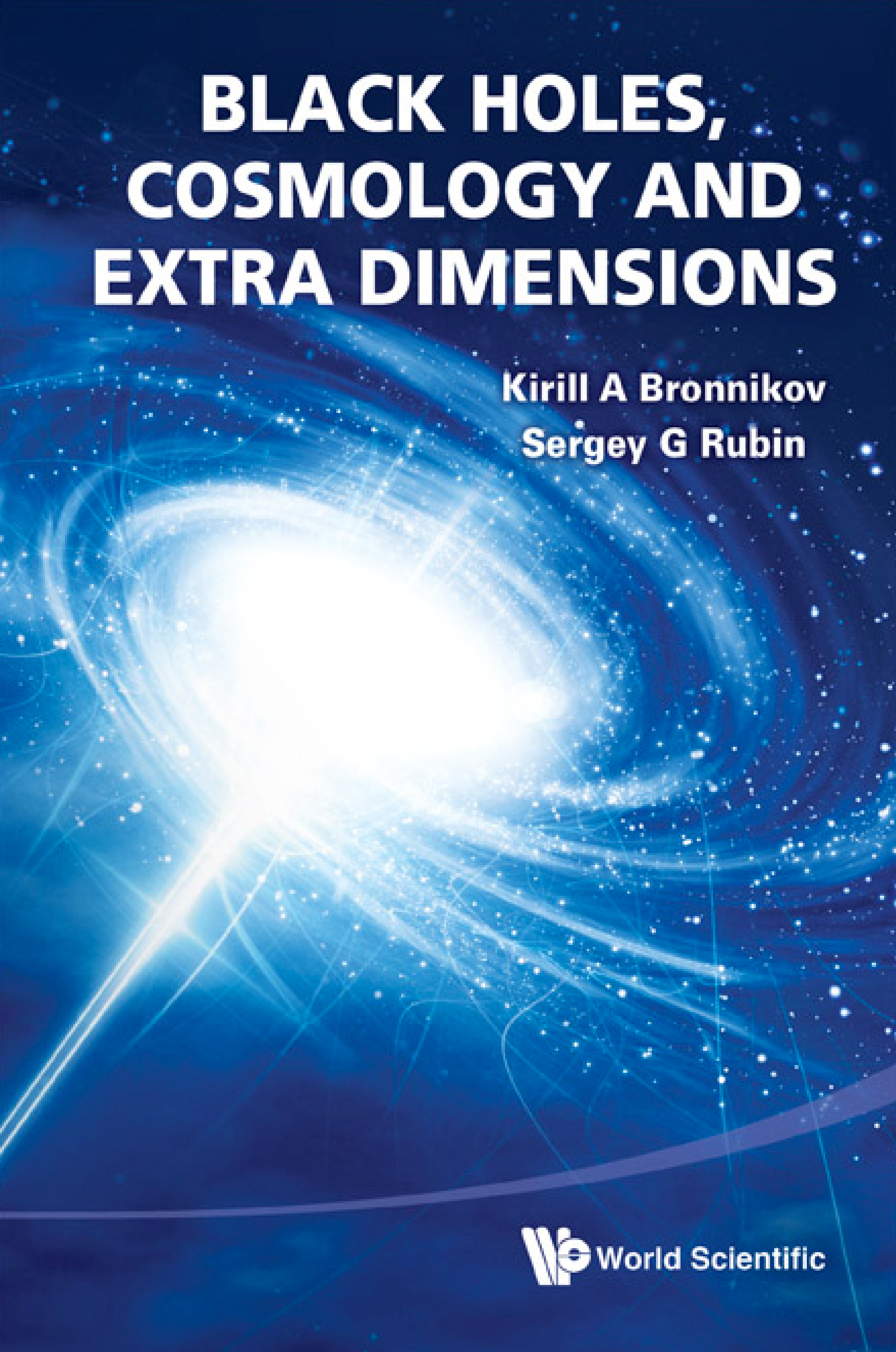
BLACK HOLES, COSMOLOGY AND EXTRA DIMENSIONS

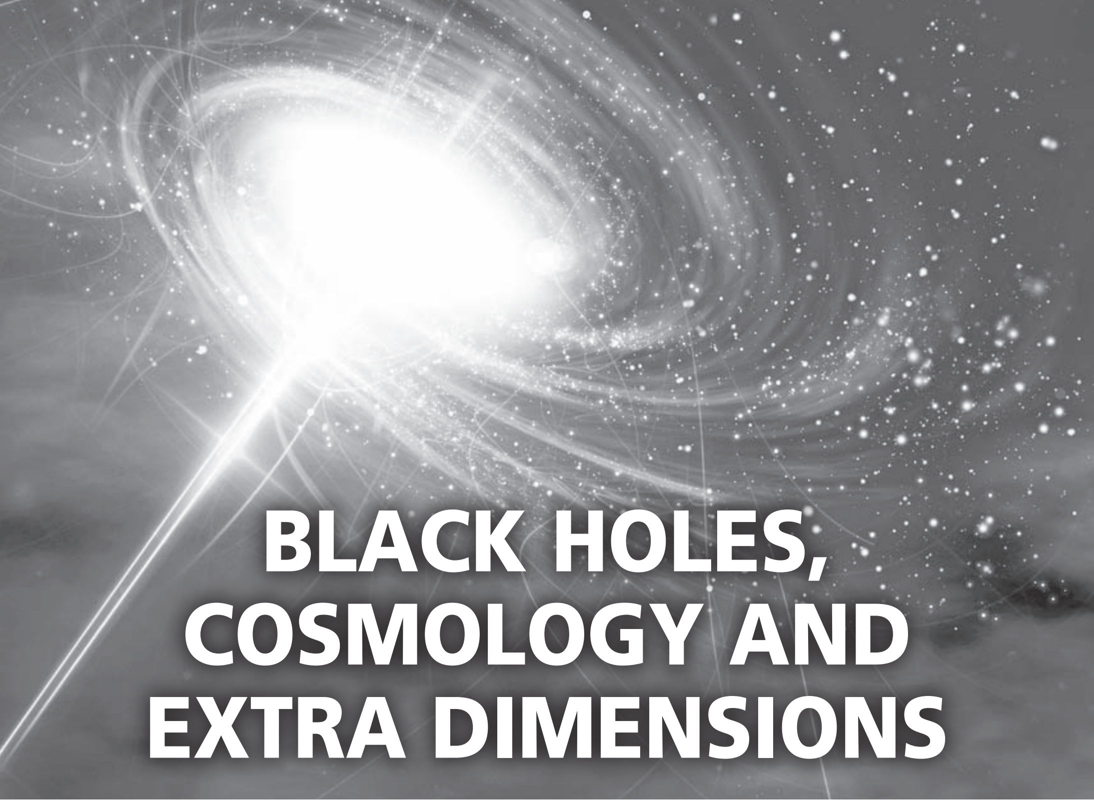
Kirill A Bronnikov
Sergey G Rubin



World Scientific

BLACK HOLES, COSMOLOGY AND EXTRA DIMENSIONS

This page intentionally left blank



BLACK HOLES, COSMOLOGY AND EXTRA DIMENSIONS

Kirill A Bronnikov
Russian Research Institute of Metrological Service, Russia

Sergey G Rubin
National Research Nuclear University "MEPhI", Russia

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

BLACK HOLES, COSMOLOGY AND EXTRA DIMENSIONS

ISBN 978-981-4374-20-0

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed in Singapore.

To our wives — Lena and Natasha

This page intentionally left blank

Contents

Notations

In the whole book, $\hbar = c = 1$ (except for Sec. 2.1);
Signature of the metric: $(1, -1, -1, -1, \ldots)$.

The cosmological metric with synchronous time:

$$ds^2 = dt^2 - \gamma_{\alpha\beta} dx^\alpha dx^\beta .$$

Units

1 light year $\sim 10^{18}$ cm.
1 parsec (pc) ~ 3 light years.
The Sun's mass: $M_\odot \sim 2 \cdot 10^{33}$ g.
The Sun's luminosity: $L_\odot \sim 4 \cdot 10^{33}$ erg/s.

The Planck units:

$M_{\rm Pl}^2 = 1/G$, or $M_{\rm Pl}^2 = 1/(8\pi G)$.
$M_{\rm Pl} \sim 10^{-5}\,{\rm g} \sim 10^{19}$ GeV.
$l_{\rm Pl} = 1/M_{\rm Pl} \sim 10^{-33}$ cm.
$t_{\rm Pl} = 1/M_{\rm Pl} \sim 10^{-44}$ s.

The main parameters of our Universe

Size: $\sim 10^{28}$ cm, or ~ 6000 Mpc.
Age: ~ 14 billion years.
The Hubble parameter: $H_0 = 100h\ {\rm km\,s^{-1}\,Mpc^{-1}}$; $h \simeq 0.72$.
Composition: $\Omega_\Lambda \simeq 0.72$; $\Omega_{\rm DM} \simeq 0.24$; $\Omega_b \simeq 0.04$.
$\sim 10^{11}$ of galaxies concentrated in clusters, with voids between them.

Our Galaxy, the Milky Way

Size: ~ 50 kpc.
$\sim 10^{11}$ of stars; neutron stars, white dwarfs, dark stars (brown dwarfs), black holes, hidden mass (dark matter).

History of the Universe, main landmarks

Time since the Big Bang	Temperature, K	Redshift	Comments
$10^{-43} - 10^{-37}$ s	$>10^{26}$		Inflation
10^{-6} s	$>10^{12}$		Quark-gluon plasma, electrons, neutrinos and their antiparticles
$3 \cdot 10^{-5}$ s	10^{12}		Baryosynthesis, nucleon formation
10^{-4} s – 3 min.	$10^{12} - 10^{9}$		D, He, Li formation
15,000 yrs	10^{4}	3000	Equal matter and radiation densities
300,000 yrs	4000	1300	Recombination, transparency of the Universe
$15 \cdot 10^{6}$ yrs	300	100	Room temperature in the Universe
$1 - 3 \cdot 10^{9}$ yrs	20	6	Birth of the first stars
$3 \cdot 10^{9}$ yrs	10	3	Formation of heavy nuclei in Supernovae
$3 - 14 \cdot 10^{9}$ yrs	3	1	Emergence of life and intelligence
10^{14} yrs	—	—	Stellar formation stops
10^{37} yrs	—	—	The last stars die
10^{100} yrs	—	—	All black holes have evaporated

Chapter 1

Modern ideas of gravitation and cosmology — a brief essay

At the beginning of the 20th century, only two physical fields were known, electromagnetic and gravitational. Einstein's theory of special relativity (SR), created in 1902–1905, described quite well the mechanical and electromagnetic phenomena at velocities up to the velocity of light, which had been impossible in the framework of Galileo and Newton's classical approach. However, Newton's theory of gravity, which was almost a perfect basis for celestial mechanics and all terrestrial physics, was formulated using the old notions of absolute space and time and could not be reconciled with SR where space and time were unified in Minkowski's four-dimensional geometry.

After the advent of SR there were numerous attempts to describe the gravitational field in Minkowski space in the hope to include rapidly moving gravitating objects into the theory. Newtonian gravity in such cases ought to be recovered in the limit of small velocities.

Henri Poincaré, great French mathematician who had actually "discovered" SR simultaneously with Einstein, was the first to try to extend it to gravity, assuming a finite propagation velocity of the gravitational field. The idea that gravity is transferred with the speed of light had been expressed before, but without SR there was no proper geometric language for such ideas to underlie a consistent theory. But the fate of special-relativistic theories of gravity was also not too happy. Despite being intrinsically free of contradiction, they still faced serious problems. One of them was their inability to explain the anomalous secular perihelion advance of planet Mercury's orbit, about $43''$ per century, which could not be explained by Newton's theory as well. This perihelion shift had been quite confidently measured in astronomical observations.

By the way, Mercury is the fastest planet of the Solar system, and it was natural to expect a manifestation of the new "gravity of high velocities" in the peculiarities of its motion.

By Kepler's laws, which follow from Newton's universal law of gravity, the planets of the Solar system move along closed elliptic orbits. This ideal picture is true if each planet interacted with the Sun only. In reality, the mutual influence of the planets upon each other slightly changes their orbits, so that their ellipses slowly rotate from one revolution to another, thus becoming unclosed spiral-like curves. These effects are calculated in Newtonian celestial mechanics with high accuracy. The anomalous perihelion shift is obtained by subtracting all "normal" shifts (explained in Newtonian physics by coordinate effects and gravitational perturbations due to other planets, whose sum amounts to about 5558 angular seconds per century) from the observed value of $5601''$ per century (see, e.g., [417] for details).

One more circumstance of theoretical nature made any attempts to describe gravity in the framework of SR, so to say, less attractive. Since Galilean times it has been well known that, if one excludes air resistance, quite different bodies — a bit of fluff, a wooden bar, a brick, a lead ingot, a can of water etc. — fall to Earth with precisely the same acceleration. The universality of free-fall acceleration was confirmed with high accuracy (up to 10^{-9}) in the Eötvös experiment with a torsion balance at the end of the 19th century: in fact, it established an equivalence between the Earth's attraction force and the inertial centrifugal force due to the Earth's diurnal rotation. In Newton's equations it is expressed as an equality between the inertial and gravitational masses, the so-called equivalence principle (EP). Newton's theory itself cannot explain this equality, as well as all its generalizations in the framework of SR.

The inertial mass that appears in the second law of Newtonian mechanics $\vec{a} = \vec{F}/m$ (acceleration is equal to force divided by mass) and the gravitational mass that appears in the law of gravity are, in essence, quantities of absolutely different physical nature. It was clear to Einstein that an equality between them cannot be a mere coincidence and should have deep reasons. The universality of the action of gravity on all bodies led him to the idea that became the basis of general relativity (GR): the gravitational field is a property of space itself, and this property should change from point to point since the gravitational field is, in general, inhomogeneous. Therefore, the Minkowski space, which is flat, homogeneous (the same at all its points) and isotropic (the same in all directions) is not suitable; gravity should warp and curve it. That is how emerges the idea of curvature of physical space-time. And since the gravitational field is created by heavy bodies, the same must be true for the curvature.

Certainly the main idea of GR, as any fundamental idea, had its forerunners and proclaimers. Even N.I. Lobachevsky, discoverer of non-Euclidean geometry, spoke in 1826 of a possible experimental determination of the world's geometry. Riemann (1854) and Clifford (1876) assumed that the curvature of space must depend on the properties of matter that fill it, and Clifford even expressed the idea that curvature may propagate by waves. So the ideas as though wandered in air. But it was Einstein (in contact with Hilbert, Poincaré and other outstanding mathematicians and physicists of that time) who converted them into an elegant and logically consistent theory.

The fruit was ripe by 1915. GR became one more step aside from the simple and clear views of classical physics: the four-dimensional space-time (or simply space, as is often said for brevity) became curved. Riemannian geometry, the geometry of curved spaces, already invented by that time, became a mathematical tool and a language of the new physical theory.

In Riemannian geometry, hence also in GR, the basic quantity used to describe the space-time is the symmetric metric tensor $g_{\mu\nu}(x^\alpha)$ (the metric), which depends on the four coordinates x^μ and consequently changes from point to point. The metric carries information on intervals between close space-time points, or events, and in its terms one expresses the quantities that directly characterize the space-time curvature, the Riemann and Ricci tensors. It is the metric tensor components that are the unknowns in the dynamic equations of GR, the Einstein equations (or the Hilbert-Einstein equations, as they are sometimes called to emphasize Hilbert's substantial contribution in GR's creation):

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -\frac{8\pi G}{c^4} T_{\mu\nu} \tag{1.1}$$

In the general case, it is a set of ten nonlinear partial differential equations with respect to ten unknown functions of four space-time coordinates. Their basic sense is a direct relationship between the space-time curvature (the left-hand side of the equations, called the Einstein tensor) and the distribution and motion of matter (the right-hand side of the equations, called the stress-energy tensor). Thus “matter tells space how to curve”. Any solution to the Einstein equations describes some possible configuration of matter and the gravitational field.

Just as SR did not cancel Newton’s mechanics (quite suitable at small velocities), GR did not cancel SR, which is valid in any small region of a curved but smooth space-time. The smaller the size of such a region, the more precisely this region coincides with a certain region of the Minkowski flat tangent space, and with greater accuracy hold SR and its numerous consequences. The Newtonian theory of gravity also follows from GR under proper conditions: Newton’s equations are obtained from those of GR in the limit of small curvature (that is, in weak gravitational fields) and small relative velocities of the gravitating bodies. Most of the observable phenomena belong to this “weak regime”. But GR interprets the gravitational forces in quite a different manner: these are now not forces but certain geometric characteristics of the world lines along which the bodies move in four-dimensional space-time. From the viewpoint of GR, a body falling freely in a gravitational field moves without any external forces, and its world line is a geodesic, a direct analogue of a straight line in flat space.

A phenomenon of utmost importance, absent in Newton’s theory but predicted by GR, is the gravitational waves. Their existence directly follows from the wave nature of GR equations and is confirmed (though so far only theoretically) by their numerous wave solutions. There is also indirect experimental data confirming their existence based on an analysis of pulsar dynamics: a binary pulsar loses its energy just as predicted by the general-relativistic radiation formulas.

GR readily responded to the observational challenge and surprisingly precisely explained the above-mentioned anomaly in Mercury’s motion. Another classical effect of GR available to verification is the effect of gravity on light rays, leading to light bending in the field of a celestial body. By Einstein’s calculation, light passing near the Sun should be bent by an angle of $1{,}75''$. A similar effect is obtained in Newtonian theory if one represents light as a flow of particles flying, naturally, with the speed of light. But then the calculated bending will be half as much as in GR: approximately $0{,}87''$ for a ray passing near the edge of the solar disk.

The total eclipse of the Sun on 29 May 1919 made it possible to measure this effect by comparing the photos of stars near the solar disk closed by the Moon with usual night photos of the same part of the sky. As was expected, for pictures taken during the eclipse, the stars turned out to be slightly shifted from the disk edge as compared to their night positions. The deflection angle varied in different observers' data from $1{,}61''$ to $1{,}98''$ near the disk edge, gradually diminishing in the outward direction, at an error within $30''$. So the sky confirmed the rightness of Einstein's prediction.

It was a true triumph: a theory created on the tip of a pen, swiftly won a place under the Sun. And what is even more amazing is that it generally preserves its leading position even now, withstanding the challenges of time and lots of experiments.

Einstein after Einstein

But let us not "run ahead of the engine" and return to the 20s and 30s of last century. It was the time of physics' active penetration into the micro-world and formation of a language adequate to its properties — quantum mechanics, later quantum electrodynamics and, wider, quantum field theory. The quantum theory was first built in the framework of the old, Newtonian concepts of absolute time and absolute space (nonrelativistic quantum mechanics), and it required substantial effort to extend it to the world of large velocities and high energies and to formulate its content in Minkowski's four-dimensional space-time.

The understanding of gravity as space-time curvature endowed GR with an exclusive character as compared to all the rest of physics, and this was in conflict with the feeling of unity of the material world, important for both philosophers and physicists. On the other hand, in GR itself there emerge quite a number of important problems, one of which is known as the problem of energy. The notions of energy and other conserved quantities play a very significant part in the structure of quantum theory. In flat space, one easily formulates the energy, momentum and angular moment conservation laws due to the symmetry of Minkowski space with respect to temporal and spatial translations and rotations, forming the 10-parameter Poincaré group. In curved space-time there are no such symmetries at all, and it is therefore quite difficult to define the energy and momentum of the gravitational field in GR in a consistent way.

For this and some other reasons, not all physicists have agreed with GR, and even now there are repeated attempts to build a theory of gravity in Minkowski space. Unlike the first attempts of this kind, the new authors have learned to explain the classical observed effects of GR, and gravitation in such theories is represented by a field with normal conservation laws and hopes to be quantized on equal grounds with other physical fields. According to Will's book [425], as early as in 1960 the number of such theories exceeded 25. But neither then nor afterwards did such theories give rise to really substantial interest (though certainly their followers would not agree with such a conclusion).

Going the other way, the contrary trend "to reduce all physics to geometry" created a number of new ideas which even now remain topical in theoretical physics. In this connection, GR was considered (and is now considered) as a basis for extension which can be achieved by introducing more complicated geometries than the Riemannian one (Weyl, Eddington, Cartan), space-time dimension larger than four with additional invisible coordinates (Kaluza, Klein), and new requirements to the symmetry of the initial formulation of the theory (Weyl's gauge symmetry principle). An ambitious task was then formulated, going far beyond a simple unification of the gravitational and electromagnetic fields: to obtain, from a unified field, all characteristics of the small number of elementary particles known at that time. And Einstein was not aside from all this effort, he was a leader of the program aimed at building a unified field theory on the basis of GR, and he remained that leader to the end of his life.

A description of these attempts would lead us too much away from our main subject, gravitation. Therefore we would like only refer to Heisenberg's words said in the early 60s: "It has been in essence a splendid attempt But at the time when Einstein developed his unified field theory, more and more new elementary particles were discovered, with their corresponding new fields. So there did not exist a steady empirical basis for carrying out Einstein's program, and his attempt did not lead to any convincing results." [Back translation from Russian.]

Even today the problem of creating a "theory of everything" remains a central problem of theoretical physics.

The technological breakthrough

By the end of the 1950s, physics already knew four rather than two basic interactions: the gravitational, electromagnetic, strong nuclear (due to which protons and neutrons join to form atomic nuclei) and weak nuclear (which is responsible for many particle transmutations and nuclear reactions of which the most well-known is beta decay). Among them, the gravitational interaction appeared to be something of minor importance: for particles, being much weaker than even the weak interaction, it seemed absolutely insignificant in micro-world physics. The accelerators supplied more and more new experimental data about the other three interactions, and quantum field theory in flat Minkowski space experienced rapid progress, formulating and solving various problems of particle physics. Against such a background, gravitational studies seemed to be something of extravagance. Yes, GR was recognized as a fundamental, almost philosophical theory, important for the world outlook, but its experimental basis was too poor: one effect, Mercury's perihelion advance, was checked up to 1%, and another, light bending near the Sun, up to roughly 30%. The cosmological observations could only testify to a nontrivial geometry of the Universe but could tell nothing about the validity of particular gravitational equations... Kip Thorne, at that time a student and now one of the most famous gravitational physicists, was advised by his professors not to deal with GR, which, in their opinion, was a theory very weakly connected with the rest of physics and astronomy. He did not obey such an advice and, as far as we can guess, hardly regrets now.

The situation began to change only in the late 50s and early 60s. The development of experimental technology made it possible to plan and carry out a number of new tests of gravity while astronomical observations gave more and more evidence of the existence of real sources of strong gravitational fields in space. As the number of alternative theories of gravity grew, tens of new effects were predicted along with suggestions to test them.

Amazingly it is GR that is being confirmed now with greater and greater accuracy. Very briefly, that is how the present experimental status of GR looks.

One of the fundamentals of GR, the equivalence principle, has been confirmed, using torsion balances and various test bodies, to an impressing accuracy of 10^{-12} by now [46]. It seems that the limit of experiments on Earth's surface has been reached by this result, due to inevitable atmospheric, seismic, and technological noise. The planned satellite experiment STEP (Satellite Test of the Equivalence Principle) [430] will raise the accuracy to 10^{-17}–10^{-18}. The equivalence principle is predicted by all metric theories of gravity, including GR and its numerous extensions where gravity is identified with space-time curvature.

Another effect, which is also universal and common to a large class of theories is the so-called gravitational redshift. Its essence is quite simple: a photon, moving away from a gravitating centre, loses energy, hence its wavelength grows (the photon "reddens"), while if it approaches the centre, it gains energy and becomes "more blue". In the same way a stone thrown up loses its speed but gains it while falling down. In GR this effect is also related to the clock slowing-down: the more we approach a source of gravity, the slower are clocks of the same physical nature. This effect has been checked both for photons (the experiments of Pound, Rebka and Snider with resonance photon absorption by atomic nuclei) [330, 331], and directly for clocks (deflections in precision atomic clock readings in air travels around the world) [193].

By the way it is this effect that makes GR not only an abstract theory but really a working tool. And, it should be stressed, very well working. The global positioning systems using satellites, of which the one named GPS is the most well-known and widespread, are more and more actively used in diverse military and civil applications and by lots of individuals. The system includes high-precision clocks, whose rate depends on both the satellite velocity (an effect of special relativity) and the Earth's gravitational field (an effect of GR), therefore corrections taking into account these effects are included in the signal calculation software: the on-board clocks are periodically "slowed down" adjusting them to clocks located on Earth. One satellite revolution around the Earth yields a clock difference that, if neglected, leads to an error of 50 to 100 meters in determining the coordinates of a receiver on Earth; meanwhile, the positioning accuracy is on the order of a meter.

The light (and radio signal) bending effect has been checked many times and with high accuracy, approaching 10^{-4} arcseconds. It became a basis for the theory of gravitational lensing, the basic method used for discovering otherwise-invisible heavy bodies in the Universe.

One more confirmation of GR is a measurement of radar echo delay in the solar gravitational field (the Shapiro effect, sometimes called the fourth classical effect of GR, in addition to the perihelion advance, light bending and redshift). The delay is not caused by lower velocity of the signals (the speed of light in vacuum is the same everywhere since special relativity holds in a small neighborhood of every point) but by their longer distance as compared to the calculated trajectory in flat space. The experiments consisted of active radar tracking of spacecrafts, and the most precise results have been obtained by using the Viking orbital and landing components that reached the planet Mars in 1977 [369]. The effect itself amounted to about 250 microseconds while the signals travelled for about an hour in the interplanetary space.

A measurement of a qualitatively different effect of GR at last produced its final result in 2011: the mission is called Gravity Probe B [159, 426]. It was an amazing project that took 47 years and 750 million dollars [426]. It was designed to test two fundamental predictions of GR, connected with precession of a gyroscope in the gravitational field of a rotating body (this time, the Earth) that had never been measured before, using four high precision gyroscopes on board a satellite orbiting the Earth at about 640 km above its surface. One of the effects is called geodetic precession: the space-time curvature exerts a torque on the gyroscope so that its axis slowly precesses, in the relevant case, by about 6.6 arcseconds per year, in the plane of the satellite's orbit. The other phenomenon, called frame-dragging, or the Lense-Thirring effect, causes a precession of the gyro axis perpendicular to the orbital plane and amounts to 40 milliarcseconds per year. The results once again well confirm the GR predictions: $6.602 \pm 0.018''$ vs. predicted $6.606''$ for the first effect (a relative uncertainty of about 0.27%), and $0.0372 \pm 0.0072''$ vs. predicted $0.0392''$ for the second effect (a relative uncertainty near 18%) [159].

All effects of GR in the Solar system are small corrections to the predictions of classical physics, which have been confidently checked up to 10^{-3} by now, and a number of modern projects suggest an improvement by a few orders of magnitude, reaching the second post-Newtonian approximation.

Much more interesting phenomena might be expected in strong gravitational fields. Observations bring us more and more new data on the manifestations of strong gravitational fields in the Universe. Thus, neutron stars, whose existence had been predicted by Oppenheimer and Volkoff as early as in the 1930s (by the way, on the basis of GR) [320], were discovered in 1967 as radio pulsars [202]. A rapid development of pulsar astronomy and physics has led, among other things, to new tests of GR. The pulsars are superdense objects with masses of the order of the solar mass and size of a few kilometers; quite often they are components of binary systems, sometimes rather close. In such systems there "operate" gravitational fields stronger than those of the Solar system by factors of hundreds and even thousands. The outstandingly high stability of pulsar "clocks" allow for accurate tracking of the celestial mechanics of such a binary. This, in particular, has brought about new confirmations of the GR prediction of secular pericentre (analogue of perihelion) shifts in binary systems.

An even more substantial contribution to gravitation theory from pulsar observations is related to gravitational waves. Despite enormous effort, they have not yet been observed directly, but binary pulsar observations unambiguously point at their trace: a slow decrease in the pulsar orbital period indicates an energy loss, and, as obtained from a calculation by general-relativistic formulas, such a loss can be explained with reasonable accuracy by gravitational wave emission. Nowadays there are hardly any people who doubt the reality of gravitational waves; on the contrary, discussed is a future advent of gravitational-wave astronomy as an important source of data on the most violent events that occur in the Universe. On the other hand, the above observational result immediately rules out many alternative theories which predict the same effects as GR in the Solar system but a more rapid energy loss by binary stars due to gravitational wave emission.

The most popular and exotic prediction of GR is certainly the existence of black holes. The image of huge insatiable mouths that devour everything and return nothing has become, no exaggeration, a part of human culture, from science fiction to folklore. Modern astrophysics considers black holes as quite real objects in space, emerging as a result of gravitational collapse of massive stars, while many phenomena in the cores of galaxies and quasars are well explained by the existence of supermassive black holes of billions of solar masses.

Curiously, the first exact solution of the Einstein equations, the Schwarzschild solution (1916, [365]), characterizing the static field of a gravitating centre, describes the simplest black hole. Though, the properties of the Schwarzschild solution were entirely understood only in the 50s [261, 392], and its peculiarities are sometimes a subject of controversy even now.

Since the early 60s, black hole physics has been developing as an independent line of investigations, and it has led to many interesting and unexpected results. Thus, it has been discovered that it is possible to mine energy from rotating black holes by launching small bodies in their neighborhood (thereby slightly slowing down their rotation) [325]; that black holes can be at full rights considered as thermodynamic objects with certain temperature and entropy [29]; that black holes "evaporate", emitting energy to the ambient space precisely as usual bodies heated to the corresponding temperature [197]. The Hawking evaporation process is connected with quantum particle production in the hole's classical gravitational field. The black holes of stellar and larger masses evaporate extremely slowly, and this does not affect any observable processes. Under real conditions, on the contrary, a black hole increases its mass at the expense of matter falling. But this process is of utmost importance for small black holes which could have survived since the first stages of the Universe expansion. Their evaporation gradually becomes stronger and stronger (since the black hole temperature grows as the mass decreases) and ends with an explosion.

In the third and fourth chapters of this book we introduce the basic notions of black hole physics and briefly discuss their main properties. A comprehensive presentation of this area of physics can be found, e.g., in the books [168, 413].

There is one more, no less intriguing consequence of the description of gravity in terms of space-time curvature. Indeed, if it is curved in principle, it is quite natural to assume that under certain conditions the curvature will be very strong — for example, it may lead to something like narrow bridges between different weakly curved universes, or “handles” connecting remote regions of the same Universe. Such geometric structures have been named wormholes. If they really exist, then, at least in principle, both time machines and interstellar travel become possible. It is not surprising that all this is readily employed in science fiction — nowadays the properties of wormholes and the conditions for their existence and construction are broadly discussed in special physical literature, see, e.g., Visser’s book [405] and the review [265]. We will touch upon some of these questions in Chapter 5.

To quantize or not?

As already pointed out, experiment entirely supports GR. However, the picture is not so unclouded in theory. We have previously mentioned the gravitational energy problem. Another well-known difficulty of GR is the existence of singularities which emerge in the majority of exact solutions to the Einstein equations and, in particular, they are hidden beyond black hole horizons and occur at the beginning and, in some models, at the end of the cosmological evolution as well as in modelling isolated bodies. These are, to put it simply, the points, lines or surfaces where the space-time loses smoothness and the quantities characterizing the curvature become either indefinite or infinite. Singularities may be connected with infinite matter densities or pressures, but there also exist purely geometric singularities, such as those in solutions to the Einstein equations in vacuum, in the absence of any matter. The inevitability of singularities in GR solutions has been proved under very general and reasonable conditions in a number of theorems, and this clearly indicates that GR is apparently not very precise in its description of extremely strong gravitational fields.

Unlike event horizons, i.e., black hole boundaries (distinguished but quite regular surfaces that work according to the remarkable principle "everybody can come in, nobody can go out"), singularities represent a real problem for the theory since they indicate, on the basis of the theory itself, the places where it does not work any more. Thus GR itself prompts to the necessity of going out of its own framework. How to do that? It is a question of great importance, the subject of many studies and discussion, maybe a question outside the framework of not only gravitation theory but physics as a whole.

It seems natural, for example, to try to account for quantum phenomena.

The relationship between gravity and quantum theory is a separate, long and very intricate story. On the one hand, as any wave field, gravitation should possess quantum properties at small scales. On the other hand, the mathematical procedures of quantization, working well in particle theory, cope very poorly with space-time curvature. There are a few ways of obtaining quantum versions of GR, but they lead to fundamentally different results. For this reason it is often claimed that quantum gravity must not be built on the basis of GR, but one should instead begin with formulating a deeper and more general theory unifying gravity with other interactions. It is well known that Einstein was for many years Bohr's opponent in discussions on the interpretation of quantum mechanics ("God does not play dice"). Quite possible, the nonacceptance of a probabilistic picture of the world by Einstein was related to its incompatibility with GR.

Some qualitative arguments prompt what might be expected from quantum gravity. Thus, from the three fundamental constants: the Planck constant $\hbar$ (quantum of action, the attribute inherent to quantum mechanics), the speed of light c (the equally fundamental constant of relativity, the maximum velocity of matter and interaction transfer) and the gravitational constant G (it is the same in Newton's theory and GR) one can construct a quantity with the dimension of length. It is called the Planck length ($l_{\rm Pl}$) and is equal to approximately 10^{-33} cm — it is probably this characteristic scale at which the space-time begins to display quantum properties. It is a very small quantity: it is smaller that an atomic nucleus (about 10^{-13} cm) by roughly the same factor by which such a nucleus is smaller than the globe ($\sim 10^7$ cm).

At such lengths there must be significant fluctuations of the metric and even topology. Accordingly, at such lengths one can expect a foam-like structure of space, which permanently changes, "breathes" at the Planck time rate $l_{\rm Pl}/c$. At still smaller lengths, the very notion of smooth spacetime becomes inapplicable, and instead of it, very likely, there is some discrete (point) structure. All this is certainly nothing more than hypotheses, but they have already given rise to numerous attempts of mathematical calculations and observational predictions.

The zoo of theories

To have a general idea of what concerns gravitational physicists nowadays, it makes sense to look through the program of some recent large gravitational conference. About a third of all talks presented will probably belong to classical GR, its astrophysical and cosmological applications. The mathematical tools are being refined, including the methods of solving the Einstein equations, new solutions are found and old ones are analyzed, questions of principle are discussed and new observational effects are suggested or predicted. In the experimental section, there are numerous works on gravitational wave detection and suggestion of measurements in space. There is inevitably a section or a few sections devoted to alternative theories of gravity, where the grand trend belongs to multidimensional theories and unification of interactions including gravity. (Let us note that the very word "alternative" — naturally, with respect to GR — stresses the peculiar role of GR among gravitational theories.) A quantum section is also certainly there.

Those who work on alternative or generalized theories pursue quite diverse objectives. There are attempts to overcome the difficulties of GR (such as, for instance, the gravitational energy problem) while preserving or strengthening its advantages; there are attempts to take into consideration principles and phenomena absent in GR. But probably the main point in all new theories is an approach to gravity as a constituent of a future "theory of everything" (or much more if not everything). The unified theories that include gravity, as a rule, use more complex geometric structures than 4D Riemannian geometry and new physical fields apart from the metric. Many of them employ ideas put forward as long ago as in the 1920s. Each of them reduces to GR under certain conditions or restrictions. And, as in GR, one seeks there solutions of physical interest (such as black holes, cosmologies, etc.) and observational predictions.

Let us mention some examples, by no means exhausting the whole diversity of approaches.

Scalar-tensor theories (STT). As is clear from the name, in these theories gravity is characterized, apart from the metric that determines the space-time curvature, by one or a few scalar fields. Let us present the Lagrangian of the Bergmann–Wagoner–Nordtvedt general STT with a single scalar field [36, 312, 412]:

$$L_{\rm STT} = f(\phi)R + h(\phi)g^{\mu\nu}\phi_{,\mu}\phi_{,\nu} - 2U(\phi) + L_m, \tag{1.2}$$

where R is the scalar curvature, ϕ is the scalar field, f, h, U are arbitrary functions of ϕ, and L_m is the Lagrangian of all other matter. The STT make the mathematically simplest extension of GR, which predicts, in general, the dependence of the gravitational constant on the scalar field and consequently on the position in space and time; it also predicts different values of classical gravitational effects from GR and more types of gravitational waves. The Lagrangian (1.2) reduces to that of GR, $L_{\rm GR} = (R-2\Lambda)/(16\pi G)$, if the field ϕ is constant in the whole space-time. (Here G is the gravitational constant and Λ is the cosmological constant.)

The observational data strongly constrain the choice of admissible STT. Without touching upon the motivations that led Jordan, Brans and Dicke in the 50s and 60s to formulation of the first STT [49, 219], let us note that numerous scalar fields naturally emerge in various multidimensional theories, in particular, those following from unification ideas, when reduced to four dimensions.

Curvature-nonlinear theories. Another important class of extended theories are those whose Lagrangian includes, apart from the scalar curvature R, some other functions of R (the so-called $f(R)$ theories) as well as other curvature invariants. Their distinguishing feature is that their gravitational field equations include third-order and higher derivatives of the metric tensor whereas this order is no more than second in GR. As a result, the set of solutions of these equations becomes much wider (though, together with the difficulty of finding them).

It should be noted that the advent of curvature-nonlinear terms in the gravitational Lagrangian is a direct consequence of quantum field theory in curved space since such terms inevitably emerge in the regularization and renormalization procedures necessary to give sense to the results of calculations [38, 178]. In other words, any theories which do not include curvature-nonlinear contributions, should be regarded as approximate ones, able to describe reality at small curvature only (certainly, to the extent that we trust the methods of quantum field theory in curved space-time).

Gauge theories. The basic idea of gauge theories traces back to the works of Hermann Weyl of 1918–1922, where he suggested using the equations of gravity and electromagnetism with a richer set of symmetries than the one contained in the Einstein equations. The additional symmetry includes some transformations of the fields themselves. Beginning with the 50s (the works of Yang and Mills), such symmetries, called local gauge symmetries, are widely used for a description of particles' interactions. It is this basic idea that has led to a unification of the weak nuclear interaction, responsible, e.g., for beta decay, with the familiar electromagnetic interaction (Weinberg, Salam, Glashow). There are promising versions that unify the resulting "electroweak" interaction with the strong nuclear one: such theories are called Grand Unification Theories (GUTs). But all this does not concern gravity. What is important is that gauge symmetries can be described in terms of geometries of some special spaces (fiber bundles), thus continuing the trend of geometrization of physics, see, e.g., the books [39, 128, 248].

Attempts to include gravity in the general scheme of gauge theories have led to different variants of gravitational theories with torsion which, in addition to the curvature, is one of the forms of "warping" the flat geometry (the Einstein–Cartan theory, the Poincaré gauge theory and others [166, 215, 328]). It has turned out that in such theories one can get rid of many kinds of singularities which exist in GR solutions as well as give new formulations of the problems of energy and quantization.

Multidimensional theories, models of superstring origin. One more idea related to geometrization of physics rests on the possible existence of extra space-time dimensions. It traces back to the pioneering works of T. Kaluza and O. Klein (1921) related to attempts to unify gravity and electromagnetism. In recent decades this idea has become a necessary ingredient of almost all attempts to unify all the four physical interactions.

Among the candidate "theories of everything", probably the most popular are the so-called superstring theories [176, 230]. Strings are one-dimensional micro-objects which, like, say, guitar strings can oscillate with a certain spectrum of frequencies. These frequencies correspond to energies of different particles. The prefix "super" designates the presence of the so-called supersymmetry, a symmetry between bosons (particles with integer spin) and fermions (particles with half-integer spins). Supersymmetry requires that each boson have a fermion counterpart. Superstrings "live" in curved spaces of dimensions 10 or 11 (depending on the particular version of the theory) and, under certain conditions, lead to a 10- or 11-dimensional version of GR. Classical GR is then obtained after distinguishing four "usual" coordinates in such spaces and inherits a large set of diverse fields from its multidimensional prototype.

Brane worlds, wells and trenches. Let us say a few words on one of the "fashionable" trends in the theory of gravity, the brane-world concept, related to unification ideas and superstrings but being of interest of their own as such.

All theories formulated in space-times of dimension higher than four have to answer the question of why the extra dimensions are still not observed. In the majority of models, beginning with those due to Kaluza and Klein, the answer is like that: the extra dimensions are closed, or compact, and have an extremely small size. A tube (a two-dimensional surface) of a micron in diameter cannot be distinguished from a geometric line, a one-dimensional object, without a microscope. The same happens to the fifth or tenth dimension: since our instruments have not reached a sufficiently small size (or, from the viewpoint of particle physics, a sufficiently high energy), the world for us is four-dimensional.

But another answer is also possible: the fifth dimension is not small, it can even be infinite, but ourselves and all observable particles and fields are "confined" in a single four-dimensional surface, or a thin layer, and a huge energy is required in order to leave it. It is like sitting on the bottom of a deep well having no power to climb out. The well is extended in three spatial dimensions, so it is more like a trench than a well: you can move along it where you wish, but not up or aside.

Such a distinguished surface or layer is called a brane (the cut-off word "membrane"), and the whole concept is called the "brane world". The first models of this sort were suggested in the 80s [10, 12, 322, 351] and became very popular in a few recent years. It became clear that they lead to new approaches in quite a number of problems of particle physics and cosmology. From the viewpoint of the theory of gravity, it is of interest that the gravitational field equations on the brane (the only field that we can observe) are more complicated than the Einstein equations and lead to somewhat different predictions. Thus, Newton's law of gravity should be violated at distances smaller than a fraction of a millimeter: instead of a squared distance, there should appear its higher power in the denominator (see, e.g., the review [75]). A discovery of such a modification of Newton's law would be a weighty argument in favor of extra dimensions in reality. Brane worlds can contain nonsingular black holes [60], wormholes without exotic matter [87] and other nonstandard phenomena.

It seems to be a good time to stop and look at the situation as a whole.

We have seen that, despite the splendid observational status of GR, most of the specialists do not believe that it is the last word in this area of physics, but only a low-energy limit of a so far unknown fundamental theory, most probably, a multidimensional one, unifying all interactions and free of such difficulties as singularities, the problem of energy and ambiguity of quantization.

Variable "constants". Among experimental predictions of the alternative theories of gravity there are quite a number of effects, such as the already-mentioned difference in the properties of gravitational waves. But probably the most striking new prediction as compared to GR is a common fact for most of the new theories, variability of quantities which had been regarded before as Fundamental Physical Constants (FPCs). In the first place, it is the gravitational constant G that appears both in Newton's law of gravity and in the Einstein equations. The electron charge, the elementary particle masses and many other constants can also vary, though the characteristic times of their variation should be comparable with the age of the Universe or even exceed it, otherwise such a variability would have been discovered long ago.

To date, the experiments and observations give only upper limits of such FPC variations. But there are two exceptions. First, beginning with the 70s, now and then there appear data on variations of G on the level $\sim 10^{-12}$ per year, but there is no confident result confirmed by different research groups. Second, observations of optical spectra of remote quasars [415] testify to slow variations of the fine structure constant $\alpha = e^2/(\hbar c)$, where e is the electron charge [74]. The most recent data [416] testify to spatial variations of α, more precisely, that quasar observations in the Northern hemisphere (mostly by the Keck telescope, Hawaiian islands) give slightly smaller values of α in the remote past while the opposite part of the celestial sphere shows values of α slightly larger than the modern one, α_0, measured on Earth (by approximately 10^{-5}). The anisotropy has a dipole nature and has received the name "Australian dipole" [34]. The existing data are summarized by the following expression for deflections of α from α_0 at any point of space:

$$\delta\alpha/\alpha_0 = (1.10 \pm 0.25) \times 10^{-5}\, r \cos\psi,$$

where ψ is the angle between the measurement direction and the axis of the dipole (declination -61 ± 9 degrees, right ascension 17.3 ± 0.6 hours), and the distance r is measured in billions of light years. There have already appeared theoretical models trying to explain both spatial and temporal variations of α [111, 319]. These models include a scalar field interacting with electromagnetism (thus affecting the value of α) and forming a cosmic domain wall due to the properties of its potential.

More details on the theoretical grounds for FPC variability and the experimental situation can be found in books and review papers [112, 246, 247, 380, 400, 401, 425].

Gravitation and the Universe

For gravitational theory, the application area of utmost importance is cosmology, the science on the Universe as a whole or on its part available to observation. Modern cosmology is a rapidly developing field of knowledge: the wealth of observational data is impetuously growing, and a lot of most diverse models are being developed. Some of the models will be discussed in detail in this book.

For a long time the Universe was considered as a kind of "vessel" containing different object: particles, stars, planets etc. It seemed that there is no relationship between the properties of this "vessel" and its content. The situation began to change with the advent of GR whose equations just established a relationship between matter and geometry. In cosmology, this relationship became even more evident after A.A. Friedmann's discovery that a stationary state of the Universe was unstable, so that it must either expand or contract. It inevitably followed from an analysis of the Einstein equations in a cosmological context. The expansion or contraction rate depends on the density and other properties of matter. The properties of the "box" have turned out to depend on its content. Further studies have led to the conclusion that the presently observed part of the Universe some time ago (about 14 billion years) had a size of about 10^{-27} cm or even smaller. But it is smaller than the size of an atom by 19 orders of magnitude. Such a small region certainly could not contain the whole wealth of particles making the stars. It means that the Universe and the particles were born simultaneously or almost simultaneously and undoubtedly exerted influence on each other. By now it is quite clear that the Universe is not a "vessel" able to contain anything but a complex organism whose parts are all mutually intertwined and interrelated. Everything is of importance here: the particle properties, gravitational physics, statistical physics, electrodynamics... and frequently all that at the same time.

A fundamental feature of the present epoch is the establishment of cosmology as a science. We can now answer many questions with proper grounds, we can explain a lot of observational data and theoretically predict many effects. The general path of the Universe's evolution is becoming clear, beginning with its creation and further into the future. Certainly not all details are clear, but it seems that almost all possible variants of the Universe's evolution have been named.

Gravity plays a key role not only in the physics of local objects, such as stars, galaxies, black holes, etc., but also in the properties of the Universe as a whole. Thus, the size of our Universe (more precisely, of its visible part) unambiguously depend on the value of the gravitational constant. The development of cosmology on the basis of GR is gradually establishing the features of the Standard Cosmological Model (SCM) whose basic significance begins to approach that of the Standard Model of particle physics.

The SCM on the basis of GR explains, above all, the following phenomena:

1. The homogeneity and isotropy of the Universe;
2. Fluctuations of the Cosmic Microwave Background (CMB) temperature;
3. The primordial nucleosynthesis.

One can assert that the history of cosmology as a science studying the Universe as a whole began in 1929 with E. Hubble's discovery. He found that the numerous nebulae, which looked on the photographs as small spots of unclear origin, are galaxies located millions of light years from our Milky Way Galaxy. It also turned out that these galaxies fly away from ours, leading to the conclusion that the whole Universe is expanding and hence there has been some time in the past when the Universe was born. Further studies have confirmed this conjecture. The abundance of helium and the existence of the CMB, discovered in 1965, also indicated that the primordial Universe consisted of a very hot plasma. It is the essence of the idea of the Big Bang.

The Universe as a whole is to a large extent homogeneous and isotropic, which is by itself hard to explain, but the large-scale structure, consisting of galaxies and clusters of galaxies, could only appear due to primordial energy density fluctuations. A detailed analysis has shown that only baryonic fluctuations are insufficient for modern structure formation. There should be some "invisible" component of energy density ("dark matter") in order that the gravitational forces be able to compress the domains of enhanced energy density and to launch stellar formation. The existence of dark matter has been by now practically proven, but its precise nature and its composition are still unclear. It is only clear that its total mass is larger than that of baryons, the particles that form stars, by approximately a factor of five.

A discovery of fundamental significance was made in 1998. It was discovered that our Universe is expanding not with a deceleration, as had been believed before, but with an acceleration. In the framework of GR it can only be explained by that the main part of the energy density is not concentrated in baryons or even dark matter but in some strange form of matter possessing a large negative pressure. It cannot consist of particles, otherwise it would form inhomogeneous structures, which is not confirmed by observations. This strange kind of matter, which has been called "dark energy", is distributed in space in a surprisingly homogeneous way, but its presence affects the galaxies' recession rate. The origin of dark energy is still unclear. The simplest explanation is that the minimum of the potential energy of a certain hypothetic field is not strictly zero but is a small positive constant. This very fact is not too surprising: what is strange is that this minimum is extremely close to zero. There is so far no widely accepted theory explaining this smallness.

There are now many open questions, but there also exist substantial advances. The most impressive step forward has occurred with the advent of the idea of inflation, i.e., a very rapid expansion of space at a primordial stage of the Universe evolution. According to the inflationary scenario, immediately after the end of inflation there happened abrupt heating of matter, reproducing the already well-known picture of a hot Universe which had been considered in pre-inflationary models as the initial stage. There are an enormous number of specific realizations of the inflationary scenario, and it is the matter of future studies to choose the correct model. Their common feature is that all of them predict an extremely rapid expansion of space in the framework of that or other theory of gravity.

One should bear in mind that cosmological models crucially depend on the choice of a theory of gravity. Therefore one can hope that cosmology, as greater and greater amounts of knowledge are stored, will not only help us to make clear the fate of our Universe but also to come to a better understanding of the mechanisms of gravity.

This page intentionally left blank

Part I

Gravitation

This page intentionally left blank

Chapter 2

Fundamentals of general relativity

The basic content of general relativity (GR) is very well presented in quite a number of well-known textbooks and monographs (e.g., [263, 296, 393, 413, 417]). Assuming that the reader is aware of these fundamentals, in this chapter, above all for reference purposes, we will mention only its basic facts and relations. Many geometric notions (such as those of a vector and a tensor, co- and contravariant components of vectors and tensors, contraction of tensors and so on) are supposed to be known and are used without explanations.

The space-time in GR is a four-dimensional manifold with a pseudo-Riemannian metric (the prefix "pseudo" is often omitted to say simply "Riemannian space-time"). The gravitational field is described in GR in terms of space-time curvature, which is expressed using the metric tensor and its derivatives with respect to the coordinates. Thus GR belongs to the class of metric theories of gravity [425]; it is historically the first, the simplest, and the most well-elaborated theory of this class.

We begin this presentation by recalling the basic facts of special relativity (SR), bearing in mind that in the close neighborhood of any point a Riemannian space-time coincides with its tangent flat (Minkowski) space-time, in which SR is formulated. So GR is not only a generalization of SR including gravity: SR holds locally in all cases as long as gravitational effects can be neglected.

2.1 Special relativity. Minkowski geometry

2.1.1 Geometry

The simplest example of a Riemannian geometry is the Minkowski flat space-time. The basic geometric invariant in this space is the interval between two arbitrary points — events. There is a class of preferable coordinates (the Minkowski coordinates) x^μ, of which each possible choice corresponds to a certain Cartesian inertial reference frame (IRF). The summed force applied to a body at rest in a certain IRF is equal to zero, or, in other words, such a body moves by inertia, uniformly and straightly with respect to any other IRF. In any IRF, the squared 4-dimensional "distance" (interval) between the events 1 $[x_1^\mu = (ct_1, \vec{x}_1)]$ and 2 $[x_2^\mu = (ct_2, \vec{x}_2)]$ is written in the form

$$s^2(1,2) = c^2(t_2 - t_1)^2 - (\vec{x}_2 - \vec{x}_1)^2, \tag{2.1}$$

where c is a universal constant coinciding with the propagation velocity of electromagnetic waves (light, in particular) in vacuum and called the speed of light. For close events 1 and 2, the interval (2.1) can be written as

$$ds^2 = \eta_{\mu\nu} dx^\mu dx^\nu, \quad \mu = 0, 1, 2, 3, \tag{2.2}$$

where the tensor with the covariant components

$$\eta_{\mu\nu} = \mathrm{diag}(1, -1, -1, -1) \tag{2.3}$$

is called the Minkowski metric tensor (the Minkowski metric). As usual, summing is assumed over repeated indices if one of them is covariant and the other contravariant. The matrix (2.3), together with its inverse matrix

$$\eta^{\mu\nu} = \mathrm{diag}(1, -1, -1, -1) \tag{2.4}$$

of the contravariant components of the Minkowski tensor are used for raising and lowering vector and tensor indices, so that, e.g., for an arbitrary vector $a = (a^\mu)$ we have $a^\mu = \eta^{\mu\nu} a_\nu$, $a_\mu = \eta_{\mu\nu} a^\nu$. The Minkowski tensor defines a scalar product (ab) of two arbitrary 4-vectors a^μ and b^μ as follows:

$$(ab) = \eta_{\mu\nu} a^\mu b^\nu = \eta^{\mu\nu} a_\mu b_\nu = a_\mu b^\mu = a^\mu b_\mu. \tag{2.5}$$

In particular, the squared interval is a scalar square of the vector $x_1^\mu - x_2^\mu$.

Evidently, the scalar square $a^2 = a_\mu a^\mu = g_{\mu\nu} a^\mu a^\nu$ of any 4-vector a can take a positive, negative or zero value because the matrix $\eta_{\mu\nu}$ is not positive- or negative-definite. In the case $a^2 > 0$ the 4-vector a is called timelike, if $a^2 < 0$, it is called spacelike, and if $a^2 = 0$ it is called null, light, or lightlike. These notions also refer to the vectors $x_1 - x_2$ that connect events 1 and 2 as well as intervals between events.[1]

Let us fix an arbitrary point O in Minkowski space and place there the origin of the coordinate system of a certain IRF. Then the whole set of ends of all null vectors x^μ beginning at O form a conic surface called the ***light cone*** (or ***null cone***) of the point O. Its equation follows from (2.1):

$$(x^0)^2 - \vec{x}^2 \equiv c^2 t^2 - x^2 - y^2 - z^2 = 0 \tag{2.6}$$

(where, as usual, $x^0 = ct, x^1 = x, x^2 = y, x^3 = z$, and $\vec{x}^2 = x^2 + y^2 + z^2$ is the 3-dimensional scalar square). All timelike straight lines that contain the point O are located inside the light cone (2.6) while all spacelike ones are located outside it.

The trajectories (world lines) of particles of nonzero mass are timelike at their any points. Since the speed of light in SR is the maximum velocity of body motion and signal propagation, two events may be causally related only if they are connected by a timelike or null interval, and one can use for them such notions as “earlier than” and “later than”. In such cases, one of the events is located either inside the light cone of the other or on the cone itself. The light cone of any event splits into the future and past light cones.

2.1.2 Coordinate transformations

A transition from one IRF to another is described in the simplest way if the velocity $\vec{v}$ of the system S' with respect to the system S is directed along one of the coordinate axes of the latter, for instance, along the axis Ox, i.e., in the system S the origin of the system moves according to the law

[1]In some books, e.g., [263], the Minkowski metric and coordinate systems are strangely called “Galilean”, despite the fact that, in Minkowski space, the (special) Einstein relativity principle is valid rather than the Galilean one, and the latter is only restored at small velocities. The usage of the word “Galilean” instead of “Minkowski” is still more confusing in connection with some new ideas in the theory of gravity, involving the Galilean transformations in some multidimensional space-times.

$x = vt$. In this case, the coordinate transformation (the special Lorentz transformation) that leaves the interval (2.1) invariant, has the form

$$x' = \frac{x - vt}{\sqrt{1 - v^2/c^2}}, \quad y' = y, \quad z' = z, \quad t' = \frac{t - vx/c^2}{\sqrt{1 - v^2/c^2}}, \tag{2.7}$$

where the primed coordinates belong to the IRF S'.

A general Lorentz transformation, connecting any two IRFs, includes a boost (a transition of the type (2.7) with a vector $\vec{v}$ of arbitrary direction) and an arbitrary rotation of the spatial coordinate axes. All Lorentz transformations form a six-parameter group called the Lorentz group. In addition to general Lorentz transformations, the interval (2.1) is invariant under space and time translations

$$x'^{\mu} = x^{\mu} + a^{\mu}, \quad a^{\mu} = \text{const}, \tag{2.8}$$

forming the four-parameter group of translations. Thus the complete group of *isometries* (coordinate transformations leaving invariant the metric tensor) contains ten parameters. It is called the Poincaré group.

The matrix of an arbitrary Lorentz transformation $A = (A^{\nu}_{\mu})$ has the definitive property of pseudo-orthogonality. It is this property that expresses the invariance of the Minkowski metric under such transformations. Namely, let the coordinates x^{μ} of the system S and the coordinates y^{μ} of the system S' be connected by the linear transformation

$$x^{\mu} = A^{\mu}_{\alpha} y^{\alpha} + a^{\mu}, \quad A^{\mu}_{\alpha},\ a^{\mu} = \text{const}. \tag{2.9}$$

According to (2.9), $A^{\mu}_{\alpha} = \partial x^{\mu}/\partial y^{\alpha}$. Substituting (2.9) to the expression for the interval (2.2), we obtain

$$ds^2 = \eta_{\mu\nu} A^{\mu}_{\alpha} A^{\nu}_{\beta} dy^{\alpha} dy^{\beta}. \tag{2.10}$$

The invariance of the metric $\eta_{\mu\nu}$ means that in the right-hand side of this equality there stands the expression $\eta_{\alpha\beta} dy^{\alpha} dy^{\beta}$, where the matrix $\eta_{\alpha\beta}$ is the same as $\eta_{\mu\nu}$ and has the form $\text{diag}(1, -1, -1, -1)$. The condition

$$\eta_{\mu\nu} A^{\mu}_{\alpha} A^{\nu}_{\beta} = \eta_{\alpha\beta} \tag{2.11}$$

is the condition that the matrix A is pseudo-orthogonal. The arbitrary constants a^{μ} in the transformation (2.9), forming the translation vector, are absent in the condition (2.11) and play the part of integration constants for the condition (2.11) treated as a set of differential equations for the functions $x^{\mu}(y^{\alpha})$.

2.1.3 Kinematic effects

An analysis of the Lorentz transformations leads to the most important kinematic effects of SR. Thus, any motion of a point particle at any fixed time instant can be considered to be approximately inertial, so one can introduce an IRF S' in which the particle is at rest at this time instant. Assuming, without loss of generality, that the motion occurs along the axis Ox, it is easy to find that the time increment dt' by a clock connected with the particle (and equal to ds/c) is related to the time increment dt by the clock of the "laboratory" IRF S according to

$$dt' = dt\sqrt{1 - v^2/c^2}. \tag{2.12}$$

Due to arbitrariness of choosing the axes, this formula is valid for a velocity of any direction, $\vec{v} = d\vec{x}/dt$. Consequently, for an arbitrary trajectory of motion $\vec{x}(t)$, the proper time interval τ of the particle (that is, the time elapsed according to a clock connected with the particle or any object whose size is insignificant) is determined by the relation

$$\tau(t_1, t_2) = \frac{1}{c}\int_{t_1}^{t_2} ds = \int_{t_1}^{t_2} dt\sqrt{1 - v^2(t)/c^2}, \tag{2.13}$$

if the time t_1 to t_2 has elapsed at the clock of an observer at rest.

From (2.12) and (2.13) it follows that the proper time interval of a moving object is always smaller than the time between the same events from the viewpoint of an observer at rest. It is the so-called ***Lorentzian time slowing-down***; it leads, in particular, to the famous ***twin paradox***. If one of the twins is at rest (or moves slowly) in a certain IRF while the other travels with relativistic velocities and, having completed his closed trajectory, meets his brother, then, at their meeting, their ages will be different: the traveller will be younger than the home-sitter. It could seem that if one considers the situation in the RF where the traveller is at rest, the result should be the opposite. However, a careful analysis, taking into account the fact that the traveller must have changed his IRF at least three times (at acceleration, at turning back and at final deceleration) shows that his calculated age will be smaller than that of his brother. The paradox is explained by the asymmetry of the situation: the integral (2.13) turns out to be smaller for an object (or subject in the present case) that has carried out noninertial motion.

Another effect, the ***Lorentzian length contraction***, consists in the fact that the length of a ruler or a rod has different values while measured in different IRFs. The length is maximum in the RF where the rod is at rest and contracts in the ratio $\sqrt{1-v^2/c^2}$ in an IRF moving with the velocity $\vec{v}$ parallel to the rod. This result is derived by finding the coordinates of the rod ends at the same time by the clock of a moving RF. The transverse size of a body does not change at its motion, and its volume changes in the same ratio as the length in the direction of motion. A body moving at relativistic velocities, comparable with the speed of light, becomes oblate, and in the ultrarelativistic limit $v \to c$ a spherical body acquires the shape of a disc or pancake.

The third effect is the relativity of simultaneity: in different IRFs, the spacelike hypersurfaces $t = \text{const}$ are different. Even more than that: any two events separated by a spacelike interval can be made simultaneous by choosing a suitable IRF. This circumstance is illustrated by Fig. 2.1 showing a two-dimensional section of Minkowski space-time making the (t, x) plane. In the coordinates x, t belonging to the IRF S, depicted are the axes x', t' of another IRF, S'; the lines $t = \text{const}$ represent the spatial sections (surfaces of simultaneity) of the IRF S, and the lines $t' = \text{const}$ (the x' axis and lines parallel to it) depict similar surfaces of the IRF S'.

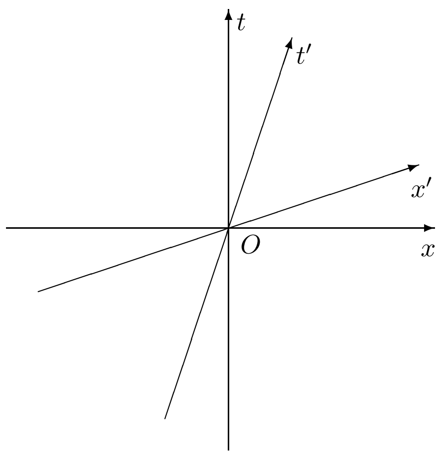


Figure 2.1 The x and t coordinate axes in different IRFs

There are other relativistic kinematic effects, e.g., light aberration (noncoincidence of directions to a distant source of light being observed from different IRFs), the relativistic formulae for velocity addition and the Doppler effect, different from their nonrelativistic counterparts etc. Their detailed analysis can be found in numerous textbooks on SR.

2.1.4 Elements of relativistic point mechanics

In the formulation of SR in terms of the Minkowski geometry, the three-dimensional quantities appearing in nonrelativistic mechanics are replaced by their four-dimensional counterparts, which leads to a more elegant, economic and transparent form of many equations. In particular, for the motion of a material point or an element of a medium, the 4-velocity ($\gamma \equiv 1/\sqrt{1-v^2/c^2}$) is introduced:

$$u^\mu = \frac{dx^\mu}{ds} = \left(\gamma, \frac{\gamma v^i}{c}\right), \tag{2.14}$$

Here $ds = c\,dt\sqrt{1-v^2/c^2}$ is an element of the interval along the particle trajectory. It is easy to verify that the vector defined in this way is normalized:

$$u_\mu u^\mu = 1. \tag{2.15}$$

The 4-acceleration of a material point or an element of a medium is

$$a^\mu = \frac{du^\mu}{ds} = \frac{\gamma}{c}\frac{du^\mu}{dt}. \tag{2.16}$$

Due to the normalization condition (2.15), the 4-velocity and 4-acceleration are mutually orthogonal: $a_\mu u^\mu = 0$.

The 4-momentum of a material point (particle) is, by definition,

$$p^\mu = m\,u^\mu = (E/c, \vec{p}), \quad m = \text{const}, \tag{2.17}$$

where m is the particle rest mass, characterizing its inertial properties, $\vec{p} = (p^i)$ is the spatial momentum, and $E = \gamma mc^2$ is the particle energy in a given IRF. By definition, the squared momentum is $p^2 = p_\mu p^\mu = m^2c^2$.

At small velocities, $v \ll c$, expanding the expression for the energy in powers of v/c, we obtain

$$E \approx mc^2 + \frac{mv^2}{2} + O(v^2/c^2). \tag{2.18}$$

From this relation follows a conclusion of utmost importance: that the total energy of a particle includes, in addition to the classical kinetic energy, the rest energy mc^2.

The role of Newton's second law is played by an equality which can be considered as a definition of the force 4-vector f^μ:

$$f^\mu = \frac{dp^\mu}{ds} = mca^\mu = mc\frac{du^\mu}{ds}, \tag{2.19}$$

where the components f^μ are connected with the usual force 3-vector by the relations

$$f^0 = \frac{\gamma}{c^2}(\vec{F}\vec{v}), \quad f^i = \frac{\gamma}{c}F^i. \tag{2.20}$$

The action of a free particle is written in the form [263]

$$S = -mc\int ds = -mc^2\int\sqrt{1-\frac{v^2}{c^2}}dt, \tag{2.21}$$

and, since by definition $S = \int L\,dt$, the Lagrangian L is

$$L = -mc^2\sqrt{1-\frac{v^2}{c^2}}. \tag{2.22}$$

For a particle moving subject to an action of external forces, this action must be supplemented with other terms which include potentials of interaction between the particle and other bodies and actions of various physical fields.

For a free particle, variation of the action (2.21) with respect to the functions $x^i(t)$, describing the particle world line, is zero, from which we obtain the Lagrangian equation

$$\frac{d}{dt}\frac{v^i}{\sqrt{1-v^2/c^2}} = 0 \Rightarrow v^i = \text{const}, \tag{2.23}$$

so that, as should be the case, a uniform rectilinear motion is obtained.

Photons and other massless particles follow lightlike (null) rectilinear trajectories and can have an arbitrary energy E and spatial momentum $p^i = (E/c^2)v^i$, so that $|\vec{p}| = E/c$. The connection between the photon energy and frequency, $E = h\nu$ appears, as is well known, not in SR but in quantum mechanics.

The relativistic mechanics satisfies the correspondence principle: Newtonian mechanics for massive bodies is restored at velocities $v \ll c$.

2.2 Riemannian space-time. Coordinate systems and reference frames

So far we have been using Minkowski coordinates in Minkowski space-time, where the metric has the form (2.2), and the inertial reference frames (IRFs) connected with these coordinates. However, nothing prevents us from describing physical processes in the framework of SR with the aid of other coordinate systems, even remaining in a fixed RF in Minkowski space. For example, one can introduce spherical or cylindrical coordinates.

In the most general cases, both in Minkowski space-time and in any Riemannian space-time, and wider, in any differentiable manifolds, arbitrary coordinate transformations $x^\mu \mapsto y^\mu$ are possible, with arbitrary functions

$$x^\mu = x^\mu(y^0, y^1, y^2, y^3). \tag{2.24}$$

In the physical space-time, the coordinate transformations (2.24) lead in general to changes in the reference frame.

One should note that a relationship between the notions of coordinate systems and reference frames is rather a subtle question which sometimes leads to confusions and misconceptions. It is therefore reasonable to explain in which sense we shall use these notions.

In this discussion, we will mostly follow the books [406, 437, 438].

To begin with, omitting a number of mathematical details, we can say that a Riemannian space (or space-time) is a differentiable manifold of arbitrary dimension D equipped with a metric $g_{\mu\nu}$ of arbitrary signature. If it is positive-definite [the signature $(+\ +\ldots+)$], the space is called proper Riemannian, in other cases it is called pseudo-Riemannian, but the prefix "pseudo" is often omitted.

We will consider an arbitrary real four-dimensional (pseudo-) Riemannian space-time with the interval

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu, \quad \mu, \nu = 0, 1, 2, 3, \tag{2.25}$$

with a symmetric metric tensor $g_{\mu\nu}$, having the signature $(+\ -\ -\ -)$.[2]

[2]Space-times with this signature are called Lorentzian. The same term is also applied to the signature $(-\ +\ +\ +)$; many authors use this signature. The physical content of the theory certainly does not depend on the choice of one plus and three minuses or vice versa, but the forms of some equations become slightly different which creates some inconvenience and requires certain care.

Fixing an arbitrary point, we can bring the metric tensor at this point to a diagonal form by linear coordinate transformations, and then, by evident additional transformations, ordering the obtained diagonal elements and normalizing them to plus or minus unity, we will always obtain the Minkowski tensor (2.3) due to our choice of the signature. This clearly indicates that in a small neighborhood of each point the geometry of any Riemannian space-time coincides with the Minkowski geometry, and all phenomena for which the curvature is insignificant can be described in the framework of SR.

2.2.1 Covariance, maps and atlases

Let us present the general definition of a coordinate system in a differentiable manifold. The definitions of all the corresponding notions and their detailed and rigorous discussions can be found in textbooks and monographs on global differential geometry, such as, e.g., [198, 270, 338].

A ***coordinate system*** in a certain region U of a differentiable manifold of class k and dimension D (or, which is the same, ***a map*** of the region U) is a one-to-one map of the region U to a certain region of the arithmetic space $\mathbb{R}^D$. The region U itself is then called the range of the map or the range of the coordinate system.

Thus each point x of the region U is put into correspondence with an ordered set of D real numbers, which numbers are called the coordinates of this point. Moreover, if U_1 and U_2 are ranges of two maps and $x \in U_1 \cap U_2$, then the coordinates of the point x in one of these maps are functions of class C^k of the coordinates of the point x in the other map, with a nonzero Jacobian of the transformation. Most frequently, manifolds of the class C^∞ are considered, with infinitely differentiable transformation functions.

The global properties of a manifold are described by sets of maps whose ranges cover the whole manifold (atlases). In other words, a union of ranges of maps belonging to a certain atlas is identical to the whole manifold.

Thus a coordinate system in a manifold and, in particular, in a four-dimensional Lorentzian space-time is simply a way to supply each point (event) with a certain "address" or label in the form of a set of numbers, and different ways of addressing must be related to each other by smooth transformations.

It is thus clear that the space-time coordinates in general do not bear any physical meaning if their values are not deliberately taken as functions of certain physical or geometric quantities (such as distances, time intervals, values of the electric field strength and so on). At the same time, the equations of a physical theory, describing objects in the space-time, contain coordinate of points as variables or parameters. But neither the properties of any physical bodies or phenomena nor any measurable quantities can depend on which labels are ascribed to the points-events. Thus social life in a city cannot depend on whether the houses are numbered by order, by meters, miles, cubits or leagues and from which point the counting begins. But the numbering can be very convenient or quite clumsy. The same with coordinates in space-time.

It then follows that *physical laws should admit a universal formulation independent of a choice of the coordinate system* (the so-called generally covariant formulation). This statement is the content of the *General Covariance Principle*.

The General Covariance Principle may be called a mathematical expression of common sense: any physical theory pretending to be reasonable should admit a generally covariant formulation. For instance, Newton's theory of gravity, special and general relativity are all generally covariant.

2.2.2 Reference frames and relativity

Let us now pass on to the notion of a *reference frame* (RF). By definition, a reference frame is an imaginary, massless, in general, arbitrarily (but smoothly) deformable body (*the reference body*), existing in a certain region of space-time and equipped at each point with perfect rulers, allowing for length measurements, and perfect clocks, allowing for time measurements.

Thus, unlike coordinate systems, which notion is purely mathematical, a RF is a physical notion, quite necessary for connecting the theory with measurements. As is well known, real prototypes of perfect clocks are atomic clocks, while real prototypes of perfect rulers are the length standards attached to the wavelengths of certain spectral lines.

The *General Relativity Principle* asserts the equivalence of all RFs in formulating the laws of nature. It is a consistent physical principle, from which follows the absence of preferred RFs in the nature. Let us note that even theories in which there are preferred RFs admit a generally covariant formulation of their equations. An evident example is SR, where inertial RFs (IRFs) play a distinguished role.

The absence of preferential RFs in the formulation of the laws of physics certainly does not mean that the results of physical measurements will be the same in all RFs: indeed, a measurement instrument is always situated in a certain RF, and this inevitably affects its readings. In particular, in SR, none of the IRFs is preferential, but measuring the length of the same rod by instruments situated in different RFs give, as we have seen, different results.

2.2.3 Reference frames and chronometric invariants

In principle, it is possible to use coordinate systems quite independently from RFs. For example, when describing the life of a city in its natural RF connected with its streets and houses, nothing and nobody forbids one from using coordinates connected with a certain system of regularly moving buses, or with the shadows of clouds floating in the sky.

In practice, however, it is much simpler and more convenient to attach the coordinates to a certain RF, assuming that the reference body is covered by a three-dimensional coordinate net x^i while the time coordinate x^0 changes along the world lines of fixed points of the reference body (they are naturally called time lines). It is said in this case that ***the coordinate system belongs to a reference frame***. The world lines of particles at rest in a given RF are described by the equation $x^i = \text{const}$ (are x^0-independent). If one makes a coordinate transformation (2.24), then the condition that the new coordinates y^μ belong to the same RF as x^μ is that the new spatial coordinates y^i are x^0-independent.

Thus the transformations

$$y^i = y^i(x^1, x^2, x^3), \tag{2.26}$$

$$y^0 = y^0(x^0, x^1, x^2, x^3), \tag{2.27}$$

are the most general transformations between coordinate systems that belong to the same RF. The equality (2.26) describes three-dimensional coordinate transformations which change the spatial coordinate net. The equality (2.27) describes arbitrary chronometric transformations that change the course of arbitrary (coordinate) clocks as well as their synchronization from one spatial point to another.

In any coordinate system x^μ, one often considers ***simulatneity surfaces, or spatial sections***, i.e., the sets of space-time points (events) defined by the condition $x^0 = \text{const}$.

It is clear from Eq. (2.27) that the condition $x^0 = \text{const}$ does not at all guarantee $y^0 = \text{const}$. In other words, even in coordinates belonging to the same RF, there can be different spatial sections, since there can be different sets of coordinate clocks (not to be confused with perfect clocks) which are synchronized in different ways and having different (and differently variable!) rates. A good example: different time scales on the Earth, i.e., manifestly in the same RF: Greenwich time, times in other time zones, the true solar time etc. Also, time according to clocks with pendulums of fixed length (whose rate depends on the gravity force which depends on latitude and altitude) and atomic time which is thought to be universal and is identified as the true physical time.

All measurable physical quantities in a given RF should be covariant under the transformations (2.26) and invariant under (2.27) (***chronometrically invariant***). This is the ***chronometric invariance principle*** introduced by Zel'manov [437]. It is dictated by the same reasoning as the general covariance principle but applied to a particular RF.

It is not hard to obtain expressions for chronometric invariants from the components of any four-dimensional tensors. Thus, any 4-vector A_μ has the following chronometrically invariant (ChI) components:

$$A^i, \quad i = 1, 2, 3; \quad A_t = A_0/\sqrt{g_{00}}. \tag{2.28}$$

Dealing with higher-rank tensors, one must do the same with each index. More specifically, from the components of an arbitrary tensor $A_{\mu_1\mu_2\dots\mu_r}$ of rank r one constructs the quantities

$$A_{0\dots0}^{i_1 i_2 \dots i_m} (g_{00})^{-(r-m)/2}, \tag{2.29}$$

where m ranges from zero to r while the number of lower indices (zeros) is equal to $r - m$. The expressions (2.28) are ChI and form contravariant tensors with respect to the spatial transformations (2.26). This result is verified in a straightforward manner. The metric coefficient g_{00} is always nonzero in coordinates belonging to a certain RF.

A very important example is the 3D chronometric metric tensor h_{ik}: according to the above, its contravariant components coincide (up to the sign) with their 4D counterparts:

$$h^{ik} = -g^{ik}. \tag{2.30}$$

It can be shown [297, 437] that the covariant components of the chronometric spatial metric tensor and its determinant are expressed in terms of 4D quantities as follows:

$$h_{ik} = -g_{ik} + \frac{g_{i0}g_{k0}}{g_{00}}, \quad h = \det(h_{ik}) = -\frac{g}{g_{00}}. \tag{2.31}$$

The tensors h^{ik} and h_{ik} can be used to raise and lower indices in various chronometric vectors and tensors.

The ChI coordinate displacement vector dx^i and the ChI (physical) time interval $dt = \sqrt{g_{00}}dx^0$ can be used to obtain the expressions for the ChI spatial length element dl and the 4D interval:

$$dl^2 = h_{ik}dx^i dx^k, \quad ds^2 = dt^2 - dl^2, \tag{2.32}$$

the latter coinciding with the interval in SR. It is also clear that the ChI 3-velocity defined as dx^i/dt automatically turns to unity in the case of a null interval, $ds = 0$. Thus the locally measured velocity of light in GR is always the same and equal to the fundamental velocity, as should be the case since a Riemannian space-time coincides at any point with its tangent Minkowski space (up to $O(dx^\mu)$).

A further development of the chronometric invariants theory includes a reformulation of all equations in terms of ChI quantities and ChI differential operators as well as a description of RFs employing their ChI characteristics: acceleration, rotation and deformation [297, 437, 438].

One should note that it is sometimes reasonable to use coordinates that do not belong to any RF, e.g., null coordinates, and in such systems it is quite possible to have $g_{00} = 0$ in the whole space-time or its certain subsets.

An even more universal description of reference frames is realized using the so-called monad formalism [298, 406, 438], where a monad is the 4-velocity field of the reference body; this formalism considers chronometric invariants as one of its "gauges".

More well-known is the tetrad (vierbein, hedron) formalism of RF description (see, e.g., [296, 297, 346]), where an RF is characterized with the aid of an orthonormalized hedron in Minkowski tangent space at each world point. Recall that, to specify an RF according to the above definition, it is sufficient to fix only one (timelike) vector at each point (i.e., the monad), while the other three vectors of a tetrad, characterizing the positions of spatial axes, are unnecessary. However, the tetrad formalism is extremely useful for calculations of observable quantities and for the description of fermions in curved space-times. More than that, some authors consider tetrads (rather than the metric) as the dynamic variables describing the gravitational field in GR and its extensions.

2.2.4 Covariance and relativity

As we have seen, the general covariance principle is nothing else but an opportunity to use any coordinates for describing the space-time and the events in it.

On the contrary, various forms of the relativity principle are important postulates of physical theories. Thus, the Galilean relativity principle that acts in Newtonian mechanics asserts that the laws of nature are independent of the choice of an IRF in Newtonian absolute space. The special relativity principle asserts quite the same but in Minkowski space, where IRFs are related by Lorentz transformations. In both cases, the theory contains a class of privileged RFs, and the transformations in question occur inside this class.

The Cartesian coordinates in Newtonian (Euclidean) space and the Minkowski coordinates in Minkowski space are privileged in two respects. They are distinguished mathematically since their coordinate lines coincide with the orbits of isometry groups of the corresponding spaces (though, this choice is not unique: spherical and cylindrical coordinates also possess this property). On the other hand, they are distinguished physically by the convenience of describing IRFs with their aid. It is well known, however, that, in many problems of classical mechanics, curvilinear coordinates are used, and physical phenomena in various noninertial RFs are studied. In SR, there is also a necessity to study accelerated RFs and to use coordinates belonging to them (e.g., uniformly accelerated frames and Rindler coordinates [344]). In such cases, quite helpful are the methods elaborated in GR and related to generally covariant formulations of all equations.

One should say that the word “relativity” that is part of the names of both SR and GR may be said to have two faces. On the one hand, this word denotes principles which declare quite an opposite property of physical laws, their absolute nature, their RF-independence; however, one applies the laws in specific RFs, i.e., relative to a chosen class of observers. On the other hand, in both SR and GR, many things that have been absolute in classical theory, become relative, i.e., RF-dependent. In SR these are simultaneity, lengths, time intervals. In GR (as well as in other metric theories of gravity) one should add to this list the finiteness or infiniteness of the spatial volume and even the topology of 3-space. In what follows we will meet specific examples of such relativity.

2.3 Riemannian space-time. Curvature

Let us present some definitions and relations, important for what follows.

Above all, the contravariant components of the metric $g^{\mu\nu}$ form a matrix reciprocal to the matrix $g_{\mu\nu}$:

$$g_{\mu\alpha}g^{\alpha\nu} = \delta_\mu^\nu, \tag{2.33}$$

where δ_μ^ν is the Kronecker symbol, equal to unity for coinciding values of the indices and to zero for noncoinciding ones. The tensors $g_{\mu\nu}$ and $g^{\mu\nu}$ are used for raising and lowering arbitrary tensor indices.

Partial derivatives of any scalar function $f(x^\mu)$ with respect to the coordinates, $\partial_\mu f$, form a covariant vector called the gradient of f. Partial derivatives of the components of a vector A_μ or A^μ in general do not form a tensor because coordinate transformations are, in general, nonlinear. To obtain a covariant form of physical equations and for many other purposes it is thus necessary to generalize the notion of a derivative, to make it a tensor. This goal is achieved by introducing the covariant derivatives

$$\begin{aligned}\nabla_\mu A_\nu &= \partial_\mu A_\nu - \Gamma^\alpha_{\mu\nu} A_\alpha;\\ \nabla_\mu A^\nu &= \partial_\mu A^\nu + \Gamma^\nu_{\mu\alpha} A^\alpha,\end{aligned} \tag{2.34}$$

where the quantities $\Gamma^\alpha_{\mu\nu}$ (they do not form a tensor!) are called the Christoffel symbols, or affine connection coefficients agreeing with the metric (the metric connection, or the metric affinity); they are expressed in terms of the metric tensor and its first-order partial derivatives:

$$\Gamma^\sigma_{\mu\nu} = \frac{1}{2} g^{\sigma\alpha}(\partial_\nu g_{\alpha\mu} + \partial_\mu g_{\alpha\nu} - \partial_\alpha g_{\mu\nu}), \tag{2.35}$$

where for the contraction $\Gamma^\alpha_{\mu\alpha}$, using (2.35), it is easy to obtain

$$\Gamma^\alpha_{\mu\alpha} = \partial_\mu(\ln\sqrt{-g}), \quad g := \det(g_{\mu\nu}). \tag{2.36}$$

The Christoffel symbols are symmetric in their lower indices, therefore, in general, there can be as many as 40 different components (ten variants of the lower pair and four values of the upper index).

One can directly verify the tensor nature for the transformations of the covariant derivatives in coordinate transformations.

For tensors of any rank, the covariant derivatives are calculated according to the methodology (2.34) applied to each upper or lower index separately. For instance, for the mixed tensor T^ν_μ we have

$$\nabla_\alpha T^\nu_\mu = \partial_\alpha T^\nu_\mu + \Gamma^\nu_{\alpha\beta} T^\beta_\mu - \Gamma^\beta_{\alpha\mu} T^\nu_\beta. \tag{2.37}$$

Due to (2.35) (in fact, since the connection agrees with the metric), the metric tensor is covariantly constant:

$$\nabla_\alpha g_{\mu\nu} = \nabla_\alpha g^{\mu\nu} = 0. \tag{2.38}$$

It then follows that the operation of covariant differentiation commutes with index raising and lowering, which is extremely convenient in transformations of complex tensor expressions.

The next point of importance is repeated application of covariant derivatives. Usual partial derivatives are known to commute with each other; the same is true for covariant derivatives of a scalar f. The first-order derivative, $\nabla_\mu f \equiv \partial_\mu f$, coincides with the partial one; the second nabla operator is already applied to a vector (gradient), but nevertheless

$$(\nabla_\alpha \nabla_\beta - \nabla_\beta \nabla_\alpha) f = 0.$$

Applied to a vector, commutation of covariant derivatives gives

$$\begin{aligned} (\nabla_\mu \nabla_\nu - \nabla_\nu \nabla_\mu) A_\rho &= R^\alpha_{\rho\mu\nu} A_\alpha; \\ (\nabla_\mu \nabla_\nu - \nabla_\nu \nabla_\mu) A^\rho &= -R^\rho_{\alpha\mu\nu} A^\alpha, \end{aligned} \tag{2.39}$$

where the quantities $R^\rho_{\alpha\mu\nu}$ are components of a tensor called the curvature tensor, or Riemann tensor (sometimes also called the Riemann-Christoffel tensor):

$$R^\sigma_{\mu\rho\nu} = \partial_\nu \Gamma^\sigma_{\mu\rho} - \partial_\rho \Gamma^\sigma_{\mu\nu} + \Gamma^\sigma_{\alpha\nu} \Gamma^\alpha_{\rho\mu} - \Gamma^\sigma_{\alpha\rho} \Gamma^\alpha_{\mu\nu}. \tag{2.40}$$

The Riemann tensor plays a central role in Riemannian geometry since it is this tensor that characterizes the distinction of a given metric from that of flat space. For the latter, all components of the Riemann tensor are zero (and, due to its tensor nature, all its components are zero, being calculated in an arbitrary coordinate system, not only in Minkowski coordinates).

It is most convenient to describe the symmetry properties of the Riemann tensor if it is represented by its covariant components, $R_{\mu\nu\rho\sigma} = g_{\mu\alpha} R^\alpha_{\nu\rho\sigma}$. By construction, it is symmetric under permutation of the first and second pair of indices and is antisymmetric within each pair:

$$R_{\mu\nu\rho\sigma} = R_{\rho\sigma\mu\nu} = -R_{\nu\mu\rho\sigma} = -R_{\mu\nu\sigma\rho}. \tag{2.41}$$

Furthermore, the components of the Riemann tensor satisfy the Ricci identity

$$R_{\mu\alpha\beta\gamma} + R_{\mu\beta\gamma\alpha} + R_{\mu\gamma\alpha\beta} = 0, \tag{2.42}$$

where the first index is the same in all terms while the other ones are subject to cyclic permutation. Due to (2.41), one can rewrite this identity with any other fixed index.

Given the symmetry of the Riemann tensor, the number of its independent components in 4-dimensional space is 20. In the general case of a D-dimensional space this number is equal to $D^2(D^2-1)/12$. In special examples of spaces possessing substantial symmetries, the number of independent components is much smaller: for example, in de Sitter space that has the maximum possible symmetry, all components of the Riemann tensor are expressed via only one constant.

Contractions of the Riemann tensor lead to the Ricci tensor $R_{\mu\nu}$ and the scalar curvature R, also called the Ricci scalar:

$$R_{\mu\nu} = R^{\alpha}_{\mu\alpha\nu}; \quad R = g^{\mu\nu} R_{\mu\nu} = R^{\alpha}_{\alpha}. \tag{2.43}$$

A direct inspection shows that the Riemann tensor, in addition to the algebraic identities (2.41) and (2.42), satisfies the differential identities

$$\nabla_{[\sigma} R_{\alpha\beta]\gamma\delta} = 0, \tag{2.44}$$

called the Bianchi identities.[3]

Their contraction by one pair of indices gives

$$\nabla_\sigma R^{\sigma}_{\alpha\beta\gamma} + \nabla_\gamma R_{\alpha\beta} - \nabla_\beta R_{\alpha\gamma} = 0, \tag{2.45}$$

while a further contraction leads to an equality of utmost importance in gravitation theory:

$$\nabla_\alpha G^{\alpha}_{\mu} = 0, \quad G^{\nu}_{\mu} := R^{\nu}_{\mu} - \frac{1}{2}\delta^{\nu}_{\mu} R. \tag{2.46}$$

The tensor G^{ν}_{μ} is called the Einstein tensor.

To conclude the section, we would like to present a formula for the invariant volume element in a Riemannian space of arbitrary dimension D: if we try to find the volume of a small parallelepiped specified by D

[3] As usual, symmetrization is meant over indices taken in parentheses and alternation over those in square brackets.

vectors dx^μ_a (a is the number of a vector, and μ is, as usual, the number of its component), the invariant volume element is

$$dV = \sqrt{|g|}|\det(dx^\mu_a)|, \quad g := \det(g_{\mu\nu}). \tag{2.47}$$

Consequently, an invariant integral of a scalar function $f(x)$ over a certain volume V is

$$\int_V \sqrt{|g|} f(x) d^D x, \quad d^D x := dx^1 dx^2 \cdots dx^D. \tag{2.48}$$

Evidently, similar formulae are valid for integration over surfaces of any dimension $d < D$ if x^μ denotes the coordinates specified on the surface while $g_{\mu\nu}$ is the internal metric of the surface, induced by the metric of ambient space.

2.4 The gravitational field action and dynamic equations

2.4.1 The Einstein equations

In GR, the dynamic variables characterizing the gravitational field are the metric tensor components $g_{\mu\nu}$. The dynamic equations of GR are derived from Hilbert's variation principle

$$\delta S = 0, \quad S = \int \frac{R - 2\Lambda}{2\varkappa} \sqrt{-g}\, d^4x + S_m, \tag{2.49}$$

where $S_m = \int L_m \sqrt{-g}\, d^4 x$ is the action of matter, i.e., substance and all fields except the gravitational field, and Λ is the cosmological constant, which is usually negligible when considering "local" configurations (up to the scale of a cluster of galaxies) but is manifestly important on the cosmological scale. The condition $\delta S = 0$ leads to the Hilbert-Einstein equations (more frequently they are simply called the Einstein equations)

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R + g_{\mu\nu}\Lambda = -\varkappa T_{\mu\nu}. \tag{2.50}$$

Here, $\varkappa = 8\pi G/c^4$ is the Einstein gravitational constant (G is the Newtonian gravitational constant), and $T_{\mu\nu}$ is the (metric) stress-energy tensor (SET) of matter:

$$T_{\mu\nu} = \frac{2}{\sqrt{-g}} \frac{\delta S_m}{\delta g^{\mu\nu}}. \tag{2.51}$$

In (2.50) there are in general ten nonlinear partial differential equations. However, first, the freedom of choosing a coordinate system makes it possible to impose four arbitrary coordinate conditions which can be formulated as equalities involving the coefficients $g_{\mu\nu}$, and there remain only six independent equations. Second, among the remaining equations there are four differential dependences related to the identities (2.46), which results in only two real dynamic equations. The other four are constraint equations which do not contain second-order time derivatives. These circumstances are of importance for all dynamic processes in GR, and above all for gravitational waves, which can have only two independent polarizations.

Due to the contracted Bianchi identities (2.46) and the Einstein equations (2.50), the SET of matter obeys four differential equalities having the meaning of conservation laws[4]

$$\nabla_\alpha T^\alpha_\mu = 0, \tag{2.52}$$

from which one can obtain the equations of motion for matter. Thus the equations of motion follow from the field equations. This circumstance is a fundamental difference between GR and the majority of classical field theories, including the Newtonian theory of gravity, where one has to postulate the equations of motion for field sources separately from the field equations.

Evidently, if the matter sources of gravity contain more than four degrees of freedom, their equations of motion are only partly restored from the Einstein equations.

2.4.2 Geodesic equations

The equations of motion for free particles in Riemannian space-times can be obtained by varying the action (2.21) (now written for a Riemannian linear element). The variation equation, as in Minkowski space, has the meaning of a trajectory (geodesic) equation, describing the extremum of the world line length between two given points. It can be written as

$$\frac{du^\alpha}{ds} + \Gamma^\alpha_{\mu\nu} u^\mu u^\nu = 0. \tag{2.53}$$

[4]In the general case, these differential equalities do not lead to integral conservation laws because, to do so, they should have the form $\partial f^\nu_\mu / dx^\nu = 0$, and it would be more precise to say that they express the SET change related to the metric change. Nevertheless, the name "conservation laws" for the equalities (2.52) has become quite common, and even more, any tensor T^ν_μ, obeying the condition (2.52), is often called conservative.

Here s is the interval which coincides with the proper time of an observer moving along the geodesic if it is timelike and the proper length along the geodesic if it is spacelike. In all cases s is a canonical parameter.[5]

Let us derive Eq. (2.53) from the conservation law (2.52). It will illustrate the possibility of deriving the equations of motion for matter from the Einstein equations.

Let us begin with the expression for the SET of a perfect fluid, which can be obtained as a natural extension of the corresponding expression from SR [263] to Riemannian spaces:

$$T_{\mu\nu} = (\varepsilon + p) u_\mu u_\nu - p\, g_{\mu\nu}, \tag{2.54}$$

where u_μ is the 4-velocity of particles of the fluid, $\varepsilon = \rho c^2$ is its energy density, and p its pressure. In particular, for dustlike matter consisting of noninteracting particles,

$$p = 0, \quad T_{\mu\nu} = \varepsilon u_\mu u_\nu. \tag{2.55}$$

The dust grains move without influence of external forces other than gravity, therefore their equations of motion coincide with those of free particles. Let us obtain it by differentiating the tensor (2.55):

$$\nabla_\alpha T^\alpha_\mu = u_\mu \nabla_\alpha (\rho u^\alpha) + \rho\, u^\alpha \nabla_\alpha u_\mu = 0. \tag{2.56}$$

Multiplying this equality by u^μ, taking into account that $u^\mu u_\mu = 1 \Rightarrow \nabla_\alpha (u^\mu u_\mu) = 0$, we obtain from (2.56) the continuity equation $\nabla_\mu (\rho u^\mu) = 0$ (its meaning is mass conservation for the dust) and the equation of motion

$$u^\mu \nabla_\mu u_\alpha = 0, \tag{2.57}$$

which can be rewritten in the form

$$u^\mu \nabla_\mu u^\alpha = 0 \Rightarrow \frac{dx^\mu}{ds} \frac{\partial u^\alpha}{\partial x^\mu} + \Gamma^\alpha_{\mu\nu} u^\mu u^\nu = 0,$$

whence finally we obtain Eq. (2.53):

$$\frac{du^\alpha}{ds} + \Gamma^\alpha_{\mu\nu} u^\mu u^\nu = 0.$$

[5] A canonical parameter is defined as the parameter along a curve with which the geodesic equation has the form (2.53) (see, e.g., [393]). Evidently, if s is a canonical parameter, then $\bar{s} = as + b$, where a and b are numbers, is also a canonical parameter; however, in gravitation theory it is reasonable to use just the parameter equal to the interval, and the freedom of its choice is restricted to constant shifts, $\bar{s} = s + b$. Usage of other, noncanonical parameters would complicate the form of Eqs. (2.53).

The geodesics can be spacelike, timelike and null, and this nature of a particular geodesic does not change along it because $\partial_\alpha(u_\mu u^\mu) = 0$.

2.4.3 The correspondence principle

Since a Riemannian space-time coincides with Minkowski space (its tangent space) in a small vicinity of each world point, the laws of SR are approximately valid at any point. It is possible to pass over to SR in the whole space in the limit of weak gravity; the weak gravity condition is formulated as the condition that the metric is only weakly deflecting from the flat metric (in any coordinates), or that the Riemann tensor is small.

It should be noted, however, that the formal transition $\varkappa \to 0$ in a solution to the Einstein equations does not generally lead to a flat metric: instead, it leads to a certain nonflat vacuum solution (that is, with $T_{\mu\nu} \equiv 0$) of the Einstein equations.

Newton's law of gravity follows from GR under the conditions of small velocities ($v \ll c$) and weak gravity (that is, the Newtonian gravitational potential, introduced in a proper way, should be small, $V \ll c^2$). In what follows, we shall verify the existence of such a transition using as an example the Schwarzschild metric, and also the validity of the relation $\varkappa = 8\pi G/c^4$.

A transition to Newtonian gravity can also be carried out under some additional conditions as the formal transition $c \to \infty$, and an expansion of the metric in powers of c^{-1} (more precisely, in powers of v/c and V/c^2) is convenient for describing the observable effects of relativistic gravity both in GR (the post-Newtonian approximation) and in other metric theories of gravity (the parametrized post-Newtonian approximation) [425].

2.5 Macroscopic matter and nongravitational fields in GR

All kinds of matter, except the gravitational field, admit a description in the framework of SR. For their description in GR (and, in general, in any theory formulated in a Riemannian space), most frequently, the so-called *minimal coupling principle* is used, according to which all equations known in SR are extended to curved space-time by replacing all partial derivatives with covariant derivatives. We note that this trick can even increase the freedom of calculations in the framework of SR, without restriction to

Minkowski coordinates, introducing curvilinear coordinates and invoking any accelerations, both translational and rotational ones.

We will present some relations valid for nongravitational matter in curved space-time according to the minimal coupling principle.

2.5.1 Perfect fluid

We have previously presented the expression (2.54) for the SET of a perfect fluid in Riemannian space-time. Using the conservation law for it, let us derive the corresponding equations of motion, the general-relativistic analogues of the continuity equation and the Euler equation.

Rewrite the tensor (2.54) in mixed components,

$$T_\mu^\nu = (\varepsilon + p)u_\mu u^\nu - p\delta_\mu^\nu, \tag{2.58}$$

apply to it the operator ∇_ν and equate the result to zero:

$$\nabla_\nu T_\mu^\nu = (\partial_\nu w)u_\mu u^\nu + w\nabla_\nu(u_\mu u^\nu) - \partial_\mu p = 0, \tag{2.59}$$

where $w := \varepsilon + p$ is the thermal function of the fluid. Contracting (2.59) with u^μ (i.e., making its projection to the direction of u^μ) and recalling that $\partial_\mu(u_\nu u^\nu) = 0$, we obtain:

$$\nabla_\nu(wu^\nu) - u^\mu \partial_\mu p = 0. \tag{2.60}$$

Let us now make a projection of (2.59) to a direction perpendicular to u^μ. Such a projection has the form $\nabla_\nu T_\mu^\nu - u^\nu u_\mu \nabla_\lambda T_\nu^\lambda = 0$. As a result, we arrive at the perfect fluid equation of motion, the general-relativistic analogue of the Euler equation

$$wu^\nu \nabla_\nu u_\mu = \partial_\mu p - u_\mu u^\nu \partial_\nu p. \tag{2.61}$$

In nonrelativistic hydrodynamics, the continuity equation is known to represent the mass conservation law. In SR and, even more in GR, the mass is not conserved, and analogues of the continuity equation are only obtained for conserved quantities such as the number of particles if one can neglect their possible production and absorption. Then one can introduce the particle number current $n^\mu = nu^\mu$, where n is the particle number density in the RF where the fluid formed by these particles is at rest, and u^μ is the 4-velocity of this fluid. The particle number conservation law (valid in the absence of their creation, annihilation and conversion) is expressed in the equality

$$\nabla_\mu(nu^\mu) = 0, \tag{2.62}$$

quite similar to the electric charge conservation law (2.70) (see below).

More general kinds of condensed matter, such as viscous fluids, are also frequently used in gravitational problems related to astrophysics and cosmology.

2.5.2 Scalar fields

For a scalar field ϕ with arbitrary self-interaction, described by a potential $V(\phi)$, if it is minimally coupled to gravity, the Lagrangian in curved space is written in precisely the same way as in Minkowski space:

$$L_s = \frac{1}{2} g^{\mu\nu} \phi_{,\mu} \phi_{,\nu} - V(\phi), c, \tag{2.63}$$

and its variation with respect to ϕ leads to an equation that generalizes the Klein-Gordon relativistic equation

$$\Box\phi + dV/d\phi = 0, \tag{2.64}$$

with the general-relativistic d'Alembert operator

$$\Box = \nabla^\alpha \nabla_\alpha = \frac{1}{\sqrt{-g}} \partial_\alpha \left(\sqrt{-g} g^{\alpha\beta} \partial_\beta \right). \tag{2.65}$$

The scalar field SET is obtained from the Lagrangian (2.63) by its variation according to (2.51):

$$T^\nu_{\mu\,s} = \phi_\mu \phi^\nu - \delta^\nu_\mu L_s. \tag{2.66}$$

More complex forms of scalar fields are also considered in the problems of gravitation and cosmology. For instance, the so-called k-essence with Lagrangians of the general form $L_s = L(\phi, X)$ (with $X = (\partial\phi)^2$) does not violate the minimal coupling principle.

Some of these Lagrangians as well as those with nonminimal coupling will be considered later.

2.5.3 The electromagnetic field

The electromagnetic (massless vector) field is characterized by the vector potential A_μ and by the strength tensor, also called the Maxwell tensor

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{2.67}$$

The electromagnetic field Lagrangian directly generalizes the corresponding flat-space expression. For a field with sources, the Lagrangian reads

$$L_{\rm e-m} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - j^\mu A_\mu, \tag{2.68}$$

where j^μ is the electric charge current density. Its variation with respect to A_μ gives a dynamic equation that corresponds to the second pair of Maxwell equations in usual electrodynamics,

$$\nabla_\nu F^{\mu\nu} \equiv \frac{1}{\sqrt{-g}}\partial_\nu(\sqrt{-g}F^{\mu\nu}) = j^\mu, \tag{2.69}$$

and due to (2.69), the electric charge conservation law automatically holds:

$$\nabla_\mu j^\mu \equiv \frac{1}{\sqrt{-g}}\partial_\mu(\sqrt{-g}j^\mu) = 0. \tag{2.70}$$

The first pair of Maxwell equations finds its analogue in the identity that follows from (2.67),

$$\nabla_\mu F_{\nu\sigma} + \nabla_\sigma F_{\mu\nu} + \nabla_\nu F_{\sigma\mu} = 0, \tag{2.71}$$

or, equivalently,

$$\nabla^*_\mu F^{\mu\nu} = 0, \tag{2.72}$$

where ${}^*F^{\mu\nu}$ is the tensor dual to $F_{\mu\nu}$:

$${}^*F^{\mu\nu} = \frac{1}{2}\frac{\varepsilon^{\mu\nu\rho\sigma}}{\sqrt{-g}}F_{\rho\sigma}. \tag{2.73}$$

The electromagnetic field SET is obtained by varying the action of the field $F_{\mu\nu}$ with respect to the metric $g^{\mu\nu}$:

$$T^\nu_{\mu\,\mathrm{e-m}} = -F_{\mu\alpha}F^{\nu\alpha} + \frac{1}{4}\delta^\nu_\mu F_{\alpha\beta}F^{\alpha\beta}. \tag{2.74}$$

The Maxwell field (2.67) does not change if one adds a gradient of any scalar function $f(x^\mu)$ to the vector potential A_μ. This property of ***gauge invariance*** is of utmost importance in electrodynamics. This property is evidently preserved if one considers Lagrangians more general than (2.68), such as $L(F)$ and $L(F,G)$, where $F := F^{\mu\nu}F_{\mu\nu}$ and $G := {}^*F^{\mu\nu}F_{\mu\nu}$. Such forms of nonlinear electrodynamics, respecting the minimal coupling principle, can also be used in gravitational problems.

In this book, we will not consider other fields, such as the spinor (spin 1/2) and massive vector fields and that with spin 3/2. The properties of such fields are discussed in detail, as well as their applications in the problems of gravity, particle physics and astrophysics, e.g., in the books [176, 232, 297].

The minimal coupling principle is not universal, and many researchers discuss nonminimal interactions between material fields and gravity. These interactions are introduced by adding such terms in the Lagrangian that contain both a material field and some invariants of the curvature (most frequently the scalar curvature), or invariants involving both the material field and the curvature (like, for instance, $A^\mu \partial_\mu R$). Thus, a nonminimal interaction between a scalar field and gravity can be introduced by adding in (2.63) the term $\xi R\phi^2$, $\xi = \text{const}$, while a nonminimal interaction between the electromagnetic field and gravity can be described by the term $\xi R\, F_{\mu\nu}F^{\mu\nu}$. In what follows, we will deal with nonminimal interactions of scalar fields and gravity in the framework of scalar-tensor and multidimensional theories of gravity.

2.6 The most symmetric spaces

2.6.1 Isometry groups and Killing vectors

Minkowski flat space-time possesses the greatest possible symmetry, expressed in the invariance of the interval with respect to the Poincaré group (G_{10}, i.e., a 10-parameter group). A general Riemannian space-time $\mathbb{V}_4$ does not possess any symmetry, while in special cases where such a symmetry does exist, it is described by an isometry group with a certain number of parameters, obviously no more than ten. This number is equal to the number of linearly independent solutions ξ_α of the Killing equation

$$\nabla_\alpha \xi_\beta + \nabla_\beta \xi_\alpha = 0. \tag{2.75}$$

The vectors ξ^μ satisfying Eq. (2.75) (the Killing vectors) possess the following property: if all points are shifted by $\xi^\alpha d\lambda$ (where $d\lambda$ is an infinitesimal quantity), all metric relations in $\mathbb{V}_4$ remain invariable. Consecutive shifts by $\xi^\alpha d\lambda$ lead to a motion of world points along an ***orbit of the Killing vector***, whose points are all mutually equivalent. (Example: on an arbitrary surface of rotation in usual 3-dimensional space, there is a Killing vector pointing along the "parallel" and describing shifts along it by small azimuthal angles, while the whole "parallel" is an orbit of this Killing vector.)

The maximally symmetric (G_{10}) metrics in $\mathbb{V}_4$ are known to be solutions to the Einstein equations in vacuum,

$$R_\mu^\nu - \delta_\mu^\nu \Lambda = 0, \tag{2.76}$$

with the cosmological constant Λ, which corresponds to the effective SET $T_\mu^\nu = \delta_\mu^\nu \Lambda/\varkappa$. These are spaces called the ***constant curvature spaces***: the space of zero constant curvature is the Minkowski space ($\Lambda = 0$), that of positive constant curvature is the de Sitter (dS) space ($\Lambda > 0$), and that of negative constant curvature is the so-called anti-de Sitter (AdS) space ($\Lambda < 0$) (even though an author named Anti-de Sitter never existed). In constant curvature spaces, the Riemann tensor is determined by a single constant (Λ) and is expressed in terms of the metric tensor in an algebraic manner:

$$R_{\mu\nu\rho\sigma} = \frac{\Lambda}{3}(g_{\mu\rho}g_{\nu\sigma} - g_{\mu\sigma}g_{\nu\rho}), \tag{2.77}$$

whence for the Ricci tensor and scalar we obtain

$$R_{\mu\nu} = \Lambda g_{\mu\nu}, \quad R = 4\Lambda. \tag{2.78}$$

Explicit forms of the de Sitter [374] and anti-de Sitter metrics will be presented a bit later.

2.6.2 Isotropic cosmology. The dS and AdS spaces

A smaller but still high symmetry (G_6) characterizes homogeneous isotropic cosmological models, to be discussed in detail later. In accordance with their name, these space-time are characterized by spatial sections satisfying the ***cosmological principle,*** that their metric properties are the same at all points (homogeneity) and all directions (isotropy). Here we will restrict ourselves to obtaining the simplest metrics, including those of spaces with symmetries higher than G_6 — these are the dS and AdS space-times.

The general form of the metric of a homogeneous and isotropic space-time, which is called the Friedmann-Robertson-Walker (FRW) metric, is

$$\begin{aligned} ds^2 &= dt^2 - a^2(t)dl^2, \\ dl^2 &= \frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2), \end{aligned} \tag{2.79}$$

where $k = 1, 0, -1$ correspond to closed (spherical), open flat and open hyperbolic models; t is the physical (proper) time measured by perfect clocks of comoving observers. It is the metric with maximally symmetric spatial sections: for $k = +1$, each spatial section $t = \text{const}$ is a 3-dimensional sphere, for $k = 0$ it is a 3-dimensional Euclidean space, and for $k = -1$ it is a 3-dimensional Lobachevsky space. The isometry group G_6 of any of them, including translations in three directions and rotations around three axes, is at the same time the isometry group of the space-time as a whole.

The function $a(t)$ describes the common evolution of all spatial lengths and is called the scale factor.

For the metric (2.79), the Einstein tensor has nonzero components G_0^0 and $G_1^1 = G_2^2 = G_3^3$ only:

$$G_0^0 = -\frac{3}{a^2}(\dot{a}^2 + k); \quad G_{\underline{i}}^{\underline{i}} = -\frac{1}{a^2}(2a\ddot{a} + \dot{a}^2 + k), \tag{2.80}$$

where the dot denotes d/dt and the underlined index means that there is no summing over it.

Due to the Einstein equations, the SET of matter, irrespective of its particular nature, has the same structure as the Einstein tensor, therefore inevitably

$$T_\mu^\nu = \mathrm{diag}(\rho, -p, -p, -p), \tag{2.81}$$

where ρ is the energy density, and p is the isotropic pressure. It is an example of how the space-time geometry prescribes the form of matter: in the FRW metric, any kind of matter has actually the perfect-fluid SET (2.81).

The $\binom{0}{0}$ and $\binom{1}{1}$ components of the Einstein equations are written in the form

$$\frac{3}{a^2}(\dot{a}^2 + k) = \varkappa\rho, \tag{2.82}$$

$$\frac{1}{a^2}(2a\ddot{a} + \dot{a}^2 + k) = -\varkappa p, \tag{2.83}$$

Eq. (2.83) is a consequence of (2.82) if one takes into account the conservation law (2.52), which gives

$$\frac{d\rho}{\rho + p} = -\frac{3\,da}{a}. \tag{2.84}$$

Thus we have a set of two independent equations for three unknown functions of time: $a(t)$, $p(t)$, and $\rho(t)$. To make the system determined, one should add one more equation, the equation of state of matter, i.e., a relation between p and ρ, which will actually determine the form of the solution.

In the case of the simplest, so-called barotropic, equation of state $p = w\rho$ where $w = \mathrm{const}$, Eq. (2.84) gives the density evolution law

$$\rho = \mathrm{const} \cdot a^{-3(1+w)}. \tag{2.85}$$

Some characteristic values of w are:

- $w = 0$ — dustlike matter, no pressure;
- $w = 1/3$ — disordered radiation, in particular, the cosmic microwave background;
- $w = 1$ — the maximally stiff matter compatible with causality, for which the speed of sound is equal to the speed of light;
- $w = -1/3$ — disordered cosmic strings;
- $w = -1$ — the cosmological constant;
- $w < -1$ — phantom matter.

In the case of dustlike matter, $w = 0$, the law (2.85) corresponds to mass conservation in a given coordinate volume (proportional to a^3) in the course of expansion or contraction.

The hypothetic phantom matter is invoked by many authors to account for the modern accelerated expansion of the Universe. Looking at (2.85), it is easy to understand that it is indeed a very strange form of matter: as the volume $\sim a^3$ grows, its density also grows instead of decreasing. Such a behavior, being extrapolated to remote future, really leads to a specific form of space-time singularity, called the Big Rip. Let us look how it happens due to Eqs. (2.82) and (2.85).

Suppose $k = 0$ (spatially flat cosmological models). Then, with (2.85), Eq. (2.82) is easily integrated to give (provided $w \neq -1$)

$$a(t) = \text{const} \cdot |t - t_0|^{2/(3+3w)}, \quad t_0 = \text{const}. \tag{2.86}$$

For $w < -1$, we thus obtain an infinite value of a at some finite time $t = t_0$, and at this time the phantom matter density grows to infinity, while all forms of "normal" matter vanish.

The case $w = -1$, the cosmological constant, is also called the case of vacuum matter. One of the reasons is that the corresponding effective SET, $\Lambda\delta_\mu^\nu/\varkappa$, is invariant under all coordinate transformations, hence any RF is comoving for such matter — this property is incompatible with any kind of massive particles which are at rest only in some particular RF. Its other important feature is that it has a constant energy density $\rho = \text{const} = \Lambda/\varkappa$ irrespective of any evolution of the metric.

In the case $w = -1$, Eq. (2.82) leads to the following solutions at different k and Λ: for $\Lambda > 0$ we obtain

$$k = 0\colon\ a(t) = a_0\, \mathrm{e}^{\pm \mathrm{H}_0 t}, \quad a_0 = \text{const}; \tag{2.87}$$

$$k = 1\colon\ a(t) = \mathrm{H}_0^{-1} \cosh[\mathrm{H}_0(t - t_0)], \quad t_0 = \text{const}; \tag{2.88}$$

$$k = -1: \;\; a(t) = \mathrm{H}_0^{-1} \sinh[\mathrm{H}_0(t - t_0)], \quad t_0 = \text{const}, \tag{2.89}$$

where $\mathrm{H}_0 := \sqrt{\Lambda/3}$. For $\Lambda < 0$, Eq. (2.82) has a solution for hyperbolic models only, $k = -1$:

$$a(t) = \mathrm{H}_0^{-1} \sin[\mathrm{H}_0(t - t_0)], \quad t_0 = \text{const}, \quad l = \sqrt{-3/\Lambda}. \tag{2.90}$$

The metrics (2.87)–(2.89) describe the de Sitter space in different coordinates *belonging to different RFs.* It is evident that even the topology of spatial sections is different: it is $\mathbb{R}^3$ for the models (2.87) and (2.89) and $\mathbb{S}^3$ (a 3-sphere) for the model (2.88). An analysis shows (see, e.g., the book by Hawking and Ellis [198]) that the dS space-time is described completely by the metric (2.88), whereas (2.87) and (2.89) only describe certain parts.

The metric (2.90) describes the AdS space.

Chapter 3

Spherically symmetric space-times. Black holes

In this chapter, we will begin considering some questions of black-hole physics. A black hole (BH) is, by definition, a space-time region where the gravitational field is so strong that no material bodies or light rays can leave this region and escape to infinity [168].

In the opinion of the majority of astrophysicists, BHs are widespread in the Universe, a great number of BHs formed in the early Universe, they form as a result of massive star evolution, and supermassive BHs (10^6 to 10^9 solar masses) are located in the central regions of many types of galaxies, including our own Milky Way Galaxy, quasars and active galactic nuclei. BHs possess many interesting and unusual properties, both theoretical and observational ones, described in detail in the books [168, 198, 296, 413] and others as well as a lot of review papers. We here only give an elementary introduction to this complicated and rapidly developing branch of physics, and we will be mostly restricted to the simplest case of spherical symmetry. Still many of the results to be presented here (e.g., the methodology of building Carter–Penrose diagrams in the general case, exact solutions with scalar fields and others) have not yet been included in standard textbooks, and some of them, to our knowledge, have only been published in journal articles and appear in a book for the first time.

3.1 Spherically symmetric gravitational fields

Spherical symmetry is a natural assumption in describing the simplest isolated bodies and island-like configurations. Spherically symmetric space-times are invariant under spatial rotations forming the isometry group $G3$.

In the general case, a spherically symmetric metric can be written in the form (see, e.g., [263])

$$ds^2 = e^{2\gamma}dt^2 - e^{2\alpha}du^2 - e^{2\beta}d\Omega^2, \quad d\Omega^2 = d\theta^2 + \sin^2\theta\, d\phi^2, \qquad (3.1)$$

where α, β, γ are, generally, functions of the radial coordinate u and the time coordinate t. We will also use the notation $r \equiv e^{\beta}$; thus r is the radius of a coordinate sphere $u = \text{const}$, $t = \text{const}$, or the Schwarzschild radial coordinate. (Let us note from the very beginning that in curved space the spherical radius r has nothing to do with a distance to the centre, and there can even be no centre at all.) In the expression (3.1), there is a freedom in choosing a reference frame (RF): different RFs correspond to reference bodies with different radial velocity distributions with respect to each other.

In the case of a static space-time (the isometry group G_4, since rotations are supplemented with time translations), one can always choose the RF in such a way that α, β, γ depend on u only. But there still remains the opportunity of replacing the radial coordinate u with another one by transformations of the form $u = u(u_{\text{new}})$; one can fix the choice of the radial coordinate by postulating a relation between the functions α, β, γ.

In solving different problems, different variants of such coordinate conditions can be convenient. Let us enumerate some of them.

1. $e^{\alpha} \equiv 1$, $du = dl$ — the Gaussian normal coordinates; the coordinate l is the true length along the radial direction, counted from a certain fixed sphere $l = 0$.

2. $u = r$, $\gamma = \gamma(r), \alpha = \alpha(r)$ — the curvature (Schwarzschild) coordinates: r is the curvature radius of the sphere $r = \text{const}$.

3. $e^{2\beta(u)} = e^{2\alpha(u)}u^2$ — isotropic coordinates, in which the spatial part of the metric is written in a conformally flat form:

$$dl^2 = e^{2\alpha(u)}(du^2 + u^2 d\Omega^2) = e^{2\alpha(u)}d\vec{x}^{\,2}, \qquad (3.2)$$

where, in the three-dimensional linear element, Cartesian coordinates have been introduced: $\vec{x} = (x^1, x^2, x^3)$.

4. $\alpha(u) = 2\beta(u) + \gamma(u)$ — the harmonic coordinate u is particularly convenient for solving problems with scalar fields.

5. $\alpha = -\gamma$ — the quasiglobal coordinate u. This name is connected with the fact that (as we shall see later) that it is more suitable than others for describing BH and other similar metrics at both sides of horizons.

6. $\alpha(u) = \gamma(u)$ — such u is sometimes called the "tortoise coordinate" because in many important cases the metric functions change very slowly when expressed in terms of u. In this case, the metric of the (t, u) subspace takes a conformally flat form, which simplifies a consideration of wave equations.

Other coordinate conditions can also be used.

Let us present the form of some quantities for the metric (3.1) with arbitrary coordinates u and t, without specifying their choice (and even the choice of a RF) from the very beginning. To obtain them in specific coordinates like those enumerated above, one should simply substitute the corresponding coordinate condition.

For a static metric (3.1), the Ricci tensor is diagonal; if there is a t dependence, there appears the component $R_{01} \neq 0$. Let us present its nonzero components in the general case:

$$\begin{aligned}
R_0^0 &= \mathrm{e}^{-2\gamma}[2\ddot{\beta} + \ddot{\alpha} + 2\dot{\beta}^2 + \dot{\alpha}^2 - \dot{\gamma}(2\dot{\beta} + \dot{\alpha})] \\
&\quad - \mathrm{e}^{-2\alpha}[\gamma'' + \gamma'(2\beta' + \gamma' - \alpha')]; \\
R_1^1 &= \mathrm{e}^{-2\gamma}[\ddot{\alpha} + \dot{\alpha}(2\dot{\beta} - \dot{\gamma} + \dot{\alpha})] \\
&\quad - \mathrm{e}^{-2\alpha}[2\beta'' + \gamma'' + 2\beta'^2 + \gamma'^2 - \alpha'(2\beta' + \gamma')]; \\
R_2^2 &= \mathrm{e}^{-2\beta} + \mathrm{e}^{-2\gamma}[\ddot{\beta} + \dot{\beta}(2\dot{\beta} - \dot{\gamma} + \dot{\alpha})] \\
&\quad - \mathrm{e}^{-2\alpha}[\beta'' + \beta'(2\beta' + \gamma' - \alpha')] = R_3^3; \\
R_{01} &= 2[\dot{\beta}' + \dot{\beta}\beta' - \dot{\alpha}\beta' - \dot{\beta}\gamma'].
\end{aligned} \tag{3.3}$$

Dots and primes stand for $\partial/\partial t$ and $\partial/\partial u$, respectively. The expressions of R_μ^ν for a static metric are obtained from here by putting all time derivatives equal to zero.

Among the components of the Einstein tensor, of most interest to us is the component G_1^1 since it does not contain second-order derivatives in u; for a static metric,

$$G_1^1 = -\,\mathrm{e}^{-2\beta} + \mathrm{e}^{-2\alpha}\beta'(\beta' + 2\gamma'). \tag{3.4}$$

The expressions (3.3) and (3.4) are necessary for substituting into the basic equations that determine the properties of self-gravitating systems in GR, the Einstein equations (2.50), which can be written in two equivalent forms:

$$G_\mu^\nu \equiv R_\mu^\nu - \frac{1}{2}\delta_\mu^\nu R = -\varkappa T_\mu^\nu \tag{3.5}$$

and

$$R_\mu^\nu \equiv R_\mu^\nu = -\varkappa \left(T_\mu^\nu - \frac{1}{2}\delta_\mu^\nu T \right), \tag{3.6}$$

where $T \equiv T_\alpha^\alpha$ is the trace of the stress-energy tensor (SET) of matter, and we have put $\varkappa = 8\pi G$ according to the Newtonian limit of GR.

When studying any space-time, it is above all important to know whether it is regular, which means that all curvature invariants are finite at all its points, or contains curvature singularities at which at least one such invariant is infinite. In many cases, it is most helpful to check the finiteness of the Kretschmann scalar (sometimes also called the Riemann tensor squared) $\mathcal{K} = R_{\alpha\beta\gamma\delta}R^{\alpha\beta\gamma\delta}$. For a static metric (3.1), the Kretschmann scalar is a sum of squares of all nonzero components $R_{\alpha\beta}{}^{\gamma\delta}$ of the Riemann tensor:

$$\begin{aligned}
\mathcal{K} &= 4K_1^2 + 8K_2^2 + 8K_3^2 + 4K_4^2, \\
K_1 &= \mathrm{e}^{-\alpha-\gamma}(\gamma' \, \mathrm{e}^{\gamma-\alpha})' = -R_{01}{}^{01}, \\
K_2 &= \mathrm{e}^{-2\alpha}\beta'\gamma' = -R_{02}{}^{02} = -R_{03}{}^{03}, \\
K_3 &= \mathrm{e}^{-\alpha-\beta}(\gamma' \, \mathrm{e}^{\beta-\alpha})' = -R_{12}{}^{12} = -R_{13}{}^{13}, \\
K_4 &= -\,\mathrm{e}^{-2\beta} + \mathrm{e}^{-2\alpha}\beta'^2 = -R_{23}{}^{23}.
\end{aligned} \tag{3.7}$$

This happens because in this case (and in many other important cases) the tensor $R_{\alpha\beta}{}^{\gamma\delta}$ is pairwise diagonal.

It is significant that all K_i are invariant under reparametrizations of the u coordinate, i.e., under the transformations $u \mapsto f(u)$. In other words, they behave as scalars at such transformations. The same is true for mixed components of all second-rank tensors, including R_μ^ν and G_μ^ν — that is why, when writing the Einstein equations, we prefer to deal with these components rather than $R_{\mu\nu}$, $G_{\mu\nu}$ or $R^{\mu\nu}, G^{\mu\nu}$, which are sensitive to u reparametrizations. Let us note that a reparametrization $u \mapsto f(u)$ is a special case of a purely spatial transformation of coordinates, leaving the reference frame unchanged (see Chapter 2).

Since the scalar $\mathcal{K}$ is a sum of squares, for its finiteness it is necessary and sufficient that ***all*** its components K_i, or, in other words, ***all*** nonzero components of the Riemann tensor $R^{\mu\nu}{}_{\alpha\beta}$ are finite. Thus, if the scalar $\mathcal{K}$ is finite, then ***all*** invariants that can be constructed algebraically from the Riemann tensor and the metric tensor are finite, in particular, the scalar curvature R, the "Ricci tensor squared" $R_{\alpha\beta}R^{\alpha\beta}$, the invariant $R_{\alpha\beta\gamma\delta}R^{\alpha\gamma}R^{\beta\delta}$ and so on. So, the finiteness of $\mathcal{K}$ at some space-time point means that a curvature singularity is absent at this point.

In the general case of a time-dependent metric (3.1), the situation with singularities is more involved.

It should be mentioned here that curvature singularities are not the only type of singularities that can appear in physically relevant space-times. In the most general form, a singularity is defined as a point or a set of points where geodesics terminate at a finite value of their canonical parameter. In other words, they are **places of geodesic incompleteness**. This can happen in any case where the metric loses its analyticity, for instance, where some of the metric functions behave like $(x - x_0)^a$ with fractional a in terms of a manifestly admissible coordinate x.

3.1.1 A regular centre and asymptotic flatness

A **centre** in a static, spherically symmetric space-time is, by definition, a point, line or surface where $r \equiv \mathrm{e}^{\beta} = 0$, a place where coordinate spheres are drawn to points. A centre can be regular or singular; regularity, as in any space-time point, is determined by finiteness of all K_i in the expression (3.7). It is necessary to note that there can be no centre at all in a spherically symmetric space-time; this happens if the quantity r is nonzero in the whole space-time, or at least in its static region. We will encounter such behavior very soon, while discussing the properties of the Schwarzschild geometry.

With an arbitrary u coordinate, the necessary and sufficient conditions for regularity of the metric at the centre ($r = 0$) are obtained in the form

$$\gamma = \gamma_0 + O(r^2), \quad |\beta'|\,\mathrm{e}^{-\alpha+\beta} = 1 + O(r^2), \tag{3.8}$$

where γ_0 is a constant. The second condition is obtained from the finiteness requirement of the quantity K_4 in (3.7). Its meaning is that the circumference to radius ratio should take the correct value (2π) for small circles circumscribed around the centre. This guarantees local flatness of space at the centre and the existence of a tangent space. These are properties of any regular point — but for a centre one has to introduce special regularity conditions because a centre is a singular point of the class of spherical coordinate systems used.

Asymptotic flatness. As $r \to \infty$, far from gravitational field sources, in many cases (though not always) the space-time geometry should coincide with the Minkowski geometry. Such space-times are called asymptotically flat. It means, above all, that all components of the Riemann tensor turn to zero, i.e., all $K_i \to 0$. The latter condition is, however, not sufficient: for an asymptotically flat space-time it is also necessary to have a correct circumference to radius ratio (2π) for large coordinate circles, otherwise even with a zero curvature tensor the three-dimensional space will have a funnel-like asymptotic, with a deficit or excess solid angle as compared to its standard value of 4π. Such a geometry characterizes a specific type of configurations called *global monopoles*. (A two-dimensional analogue of such three-dimensional surfaces is a conical surface whole metric is flat at all points but the top, however, the length of a circle circumscribed near the top relates to its radius by a factor different from 2π.)

Just as a centre can be lacking in a spherically symmetric space-time, a flat infinity can also be lacking; moreover, there can be no limit $r \to \infty$. We have already met such a geometry: Friedmann's closed world is spherically symmetric, and its every spatial section contains (in a given spherical coordinate system) two centres but no spatial infinity.

The necessary and sufficient conditions for the metric (3.1) to be asymptotically flat as $r \to \infty$ in terms of an arbitrary coordinate u are as follows:

$$\gamma = \gamma_\infty + O(1/r), \quad |\beta'|\,\mathrm{e}^{-\alpha+\beta} = 1 + O(1/r), \tag{3.9}$$

where γ_∞ is a constant.

In the conditions (3.8) and (3.9), as everywhere, the notation $y = O(x)$ means that either x and y are quantities of the same order ($y \sim x$) or y is much smaller than x (the latter is denoted as $y \ll x$ or $y = o(x)$).

3.2 The Reissner–Nordström–(anti-)de Sitter solution

3.2.1 Solution of the Einstein equations

Let us find an important class of exact static, spherically symmetric solutions to the Einstein equations, characterizing the gravitational fields in vacuum or in the presence of an electromagnetic field (without charges) and a cosmological constant. This class contains the metrics that have the greatest number of astrophysical applications among all spherically symmetric metrics; it will also provide us with explicit examples in our future discussion of general properties of spherically symmetric space-times, including those with BHs.

It proves to be convenient to solve the problem in the curvature coordinates, in which two independent Einstein equations can be written in the form (3.5):

$$G^0_0 + \Lambda = \mathrm{e}^{-2\alpha}\left(\frac{1}{r^2} - \frac{2\alpha'}{r}\right) - \frac{1}{r^2} + \Lambda = -\varkappa T^0_0, \qquad (3.10)$$

$$G^1_1 + \Lambda = \mathrm{e}^{-2\alpha}\left(\frac{1}{r^2} + \frac{2\gamma'}{r}\right) - \frac{1}{r^2} + \Lambda = -\varkappa T^1_1, \qquad (3.11)$$

where the prime denotes d/dr while the SET in the present case corresponds to the electromagnetic field.

The Lagrangian and SET of the electromagnetic field are

$$L_{\mathrm{e-m}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \quad T^\nu_\mu = \frac{1}{4}[-4F_{\mu\alpha}F^{\nu\alpha} + \delta^\nu_\mu F_{\alpha\beta}F^{\alpha\beta}]. \qquad (3.12)$$

The Maxwell equations $\nabla_\alpha F^{\alpha\beta} = 0$ must now be written for the spherically symmetric case, so that among the components of $F_{\mu\nu}$ only the ones describing a radial electric field ($F_{01} = -F_{10}$) and a radial magnetic field ($F_{23} = -F_{32}$) can be nonzero.[6] Let us restrict ourselves to an electric field. Then the only nontrivial Maxwell equation yields

$$(\sqrt{-g}F^{01})' = 0 \Rightarrow F^{01} = \frac{Q\,\mathrm{e}^{-\alpha-\gamma}}{\sqrt{4\pi}r^2}, \quad F_{10} = \frac{Q\,\mathrm{e}^{\alpha+\gamma}}{\sqrt{4\pi}r^2}, \qquad (3.13)$$

where the constant Q is interpreted as an electric charge (we are using the Heaviside units in electrodynamic). The SET takes the form

$$T^\nu_\mu = \frac{Q^2}{8\pi r^4}\,\mathrm{diag}(+1, +1, -1, -1). \qquad (3.14)$$

Let us now solve the Einstein equations. Eq. (3.10) can be brought to the form

$$\begin{aligned}[r(A-1)]' &= -(\Lambda + \varkappa T^0_0)r^2 \\ &\Rightarrow A = 1 - \frac{\Lambda r^2}{3} - \frac{G}{r}\int T^0_0 r^2 dr, \end{aligned} \qquad (3.15)$$

where $A(r) = \mathrm{e}^{-2\alpha}$ and we have recalled that $\varkappa = 8\pi G$. On the other hand, the difference of Eqs. (3.10) and (3.11), under the condition $T^0_0 = T^1_1$,

[6]It is often claimed that a radial magnetic field is only possible if one assumes the existence of magnetic monopoles. However, in configurations like wormholes both electric and magnetic radial fields can exist without any charges, see Chapter 5.

gives

$$\alpha' + \gamma' = 0 \Rightarrow \mathrm{e}^{\alpha+\gamma} = 1, \tag{3.16}$$

under a proper choice of scale along the time axis. As a result of integration in (3.15), the metric finally takes the form

$$ds^2 = A(r)dt^2 - \frac{dr^2}{A(r)} - r^2 d\Omega^2,$$
$$A(r) = 1 - \frac{\Lambda}{3}r^2 - \frac{2GM}{r} + \frac{GQ^2}{r^2}, \tag{3.17}$$

where the integration constant M is interpreted as the active gravitational mass of the gravitational field source.

3.2.2 Special cases

The (anti-)de Sitter metric

In the special case of zero mass and charge, the function $A(r)$ in Eq. (3.17) takes the simple form $A = 1 - (\Lambda/3)r^2$. At $\Lambda = 0$, the Minkowski metric is reproduced. At $\Lambda > 0$, it can be shown that the metric describes the de Sitter space-time, already discussed in the previous chapter in a cosmological context. But here we have its static representation, hence we obtain it here in a RF different from the previous ones. At $\Lambda < 0$ we accordingly obtain the AdS space-time. Explicit relations between the static and cosmological coordinates in de Sitter space-time will be presented later.

Consider the case $\Lambda = 3H^2 > 0$ in more detail. The metric is

$$ds^2 = (1 - H^2r^2)dt^2 - (1 - H^2r^2)^{-1}dr^2 - r^2 d\Omega^2. \tag{3.18}$$

At $r = 0$ it has a regular centre. It must be so because de Sitter space-time, being a constant-curvature space, is homogeneous; all its points are equivalent and regular, and any of them can be “appointed” to be a centre.

At $r = 1/H$ the metric coefficients g_{00} and g_{11} turn to zero (this sphere is called a horizon), while at $r > 1/H$ they take negative values, so that the coordinates r and t exchange their roles: the quantity r now becomes a temporal coordinate and t a spatial coordinate. Such a region is called a T-region; see a more general description of horizons, R- and T-regions in the next section. In this T-region, the metric can be rewritten as follows:

$$ds^2 = (H^2r^2 - 1)^{-1}dr^2 - (H^2r^2 - 1)dx^2 - r^2 d\Omega^2$$
$$= d\tau^2 - \sinh^2(H\tau)dx^2 - H^{-2}\cosh^2(H\tau)d\Omega^2, \tag{3.19}$$

where the “former time” t is renamed x while the variable τ, introduced instead of r, has the meaning of proper time in this cosmological model, which is homogeneous but anisotropic. The homogeneity follows from the fact that, in addition to spherical symmetry, the metric coefficients are independent of the spatial coordinate x; the anisotropy manifests itself by the existence of two different scale factors, $|g_{xx}| = \sinh^2(H\tau)$ and $|g_{\theta\theta}| = r^2 = \cosh^2(H\tau)$, characterizing the expansion of the Universe in the x direction and in the two angular directions. Such spherically symmetric cosmological models, with the topology of the spatial section $\mathbb{R}\times\mathbb{S}^2$, whose special case is the model (3.19), are called Kantowski–Sachs models. Looking ahead, we note that Kantowski–Sachs models describe T-regions of any spherically symmetric BHs.

In the model (3.19), the cosmological expansion begins with the horizon $\tau=0$, which is an extremely anisotropic state because $g_{xx}=0$, but becomes isotropic and exponential (with the Hubble constant H) at large τ. It is one more face of the de Sitter metric in addition to the three isotropic cosmological models and (3.18).

Let us stress that in this description the metrics (3.18) and (3.19) must be considered separately from each other because at $r = H^{-1}$, $A=0$, so both metrics possess coordinate singularities. The same is true for all R- and T-regions up to the end of this section. At all horizons, corresponding to regular zeros of the function $A(r)$, as is easily verified using Eqs. (3.7), the four-metric is regular, and it is natural to suppose that a transition to new coordinates can allow for avoidance of coordinate singularities and obtaining a complete space-time picture naturally including R- and T-regions. Such transformations will be described in the next sections.

The Schwarzschild metric and the Newton law

In the case $\Lambda = 0$, $Q = 0$, the metric (3.17) turns into the Schwarzschild metric

$$ds^2 = \left(1 - \frac{2GM}{r}\right) dt^2 - \left(1 - \frac{2GM}{r}\right)^{-1} dr^2 - r^2 d\Omega^2, \qquad (3.20)$$

where the integration constant M has the meaning of an active gravitational mass. To confirm this interpretation, let us use the geodesic equation (2.53) for the case of a test particle instantaneously at rest, at a large (compared to $2GM$) value of the radius r:

$$\frac{d^2x^i}{ds^2} + \Gamma^i_{00}(u^0)^2 = 0,$$

where $ds^2 = g_{00}dt^2 \approx dt^2$ and $u^0 \approx 1$ due to the assumptions made. Among the Christoffel symbols Γ^i_{00} only one is nonzero, namely, $\Gamma^1_{00} = \gamma' \, e^{2\gamma - 2\alpha}$ (in terms of the metric (3.1)), or, with (3.20), $\Gamma^1_{00} \approx GM/r^2$. Thus the particle experiences the acceleration $-GM/r^2$ in the direction of smaller radii, as was required.

The Schwarzschild metric describes the gravitational field in vacuum around any isolated spherically symmetric body provided that the space-time far from the body can be regarded as asymptotically flat — and this is the case for an overwhelming majority of phenomena of astrophysical interest. The properties of the Schwarzschild metric will be discussed in detail in section 3.4. We will only mention here that, like the de Sitter metric (3.18), the Schwarzschild metric (3.20) contains a single horizon at $r = r_g = 2GM$, beyond which, at $r < r_g$, there is a T-region representing a certain Kantowski–Sachs model.

The region $r \le r_g$ is the simplest example of a black hole.

The Reissner–Nordström metric

The Reissner–Nordström metric is obtained from (3.17) at $\Lambda = 0$ and describes the external gravitational field of a body of mass M having the electric charge Q. The geometric and hence physical properties of the Reissner–Nordström space-time depend on the charge to mass ratio. Let us introduce the "geometrized" mass and charge having the dimension of length:

$$m = GM, \quad q = \sqrt{G}Q, \tag{3.21}$$

then under the condition $m^2 > q^2$ the quadratic trinomial (see (3.17))

$$r^2 A(r) = r^2 - 2mr + q^2, \tag{3.22}$$

has two positive roots

$$r_\pm = m \pm \sqrt{m^2 - q^2}, \tag{3.23}$$

so that $A(r) > 0$ (the metric is static) at $r_+ < r < \infty$ — it is the outer R-region, and at $0 < r < r_-$ — the inner R-region. The intermediate values $r_- < r < r_+$ correspond to a T-region, in which, as in the Schwarzschild and de Sitter space-times, the metric describes a certain homogeneous anisotropic Kantowski–Sachs model. The spheres $r = r_+$ and $r = r_-$, separating the R- and T-regions, are called the outer and inner horizons of a Reissner–Nordström BH, respectively.

Under the condition $m^2 = q^2$, so that $A = (r - m)^2$, the two horizons merge into a single one, $r = r_h = m = |q|$; it is called a **double, or extremal horizon**, and the region $r \leq m$ (or sometimes the whole configuration) is called an extremal Reissner–Nordström BH. At both sides of the sphere $r = r_h$, the function $A(r)$ is positive.

Lastly, for "large charges", $q^2 > m^2$, the whole space $r > 0$ is occupied by a single R-region.

At $r \to \infty$, the contribution of the charge to the function $A(r)$ can be neglected, and the metric approximately coincides with Schwarzschild's.

On the contrary, at small radii, $r \to 0$, the properties of the metric are completely determined by the charge. Since it is in all cases an R-region, the value $r = 0$ corresponds to a centre, and, as shown by comparison with Eqs. (3.7), this centre is singular. Since $g_{00} = A \to \infty$ as $r \to 0$, it is easy to find that the centre repels neutral test particles.

Metrics with a nonzero cosmological constant

At $\Lambda \neq 0$, the basic properties of the metric (3.17) are again determined by the behavior of the function $A(r)$ and, above all, by the number and disposition of its zeros, each of them corresponding to a horizon that separates R- and T-regions.

At $Q = 0$, $M > 0$ and $\Lambda > 0$ (the Schwarzschild-de Sitter metric), the number of horizons can range from zero to two, but at small and large r, at $r \to 0$ and $r \to \infty$, there are T-regions. If there is, in addition, a nonzero charge, then there is an R-region at $r \to 0$, with a singular centre just as in the Reissner–Nordström solution, while at $r \to \infty$ the metric behaves as in the de Sitter solution (a T-region), so that the number of horizons varies from 1 to 3 depending on the particular values of the three parameters m, q and Λ.

3.3 Horizons and geodesics in static, spherically symmetric space-times

After the first appearance of horizons in the de Sitter, Schwarzschild and Reissner–Nordström metrics and their generalizations, it makes sense to make the notion of a horizon more exact and to establish its close relationship with the notion of a black hole (BH). As was mentioned at the beginning of this chapter, a BH is by definition such a space-time region that no material body or light signal can leave it. Since the motion of massive and massless particles is described in metric theories of gravity by geodesic equations, the question of which space-time region is a BH is answered by studying the properties of geodesics.

It is for this reason that this section begins with a description of the general form of geodesics in an arbitrary static, spherically symmetric space-time; after that, we will discuss the behavior of geodesics in the vicinity of horizons and their relationship with the BH notion. We will conclude the section by briefly describing different kinds of horizons mentioned in papers on the theory of gravity.

3.3.1 The general form of geodesic equations

Let us return to the general static, spherically symmetric metric (3.1) with an arbitrary radial coordinate u and consider the geodesic equations (2.53) as equations with respect to the unknown functions $x^0 = t(\lambda)$, $x^1 = u(\lambda)$, $x^2 = \theta(\lambda)$, $x^3 = \phi(\lambda)$, the coordinates of a point on the trajectory, as functions of the canonical parameter λ.

For simplicity, and in connection with the symmetry of the problem, we will assume that the geodesic is located in the equatorial plane $\theta = \pi/2$; we denote $d/d\lambda$ by a dot and d/du by a prime. The equations, along with first integrals of two of them, have the form

$$\ddot{t} + 2\gamma'\dot{t}\dot{u} = 0 \Rightarrow \dot{t} = E\,\mathrm{e}^{-2\gamma}, \tag{3.24}$$

$$\ddot{u} + \gamma'\,\mathrm{e}^{2\gamma-2\alpha}\dot{t}^2 + \alpha'\dot{u}^2 - \beta'\,\mathrm{e}^{2\beta-2\alpha}\dot{\phi}^2 = 0, \tag{3.25}$$

$$\ddot{\phi} + 2\beta'\dot{\phi}\dot{u} = 0 \Rightarrow |\dot{\phi}| = L\,\mathrm{e}^{-2\beta}, \tag{3.26}$$

where E and L are integration constants. The equation that contains $\ddot{\theta}$ is trivial.

Eqs. (3.24)–(3.26) are not independent: there is the constraint $u_\alpha u^\alpha = k$, where $u^\mu = dx^\mu/d\lambda$; $k = +1$ for timelike geodesics (in this case u^μ is the 4-velocity and the parameter λ coincides with the proper time), $k = 0$ for null geodesics, and $k = -1$ for spacelike ones. Let us write down the explicit form of this constraint, which represents an integral of Eq. (3.25):

$$\mathrm{e}^{2\gamma}\dot{t}^2 - \mathrm{e}^{2\alpha}\dot{u}^2 - \mathrm{e}^{2\beta}\dot{\phi}^2 = k, \tag{3.27}$$

and substitute the integrals (3.24) and (3.26). After multiplying by $\mathrm{e}^{2\gamma}$ we obtain

$$\mathrm{e}^{2\alpha+2\gamma}\dot{u}^2 + k\,\mathrm{e}^{2\gamma} + L^2\,\mathrm{e}^{2\gamma-2\beta} = E^2. \tag{3.28}$$

The latter relation has the form of an energy conservation law for a particle moving in a potential field along the u axis; the quantity $E^2 \geq 0$ plays the part of the total energy, the first term is an analogue of the particle's kinetic energy while a sum of the second and third ones makes an analogue of the potential energy; moreover, the effective potential

$$V(u) = e^{2\gamma}(k + L^2 e^{-2\beta}) \tag{3.29}$$

plays the same role for geodesic motion as the potential in classical mechanics for a one-dimensional motion of a point particle: the motion is only possible in a region where $E^2 \geq V(u)$, while the values of the coordinate u at which $E^2 = V$ correspond to turning points.

The constant L related to changes in the azimuthal angle ϕ can be interpreted as the particle's conserved angular momentum.

3.3.2 Horizons, geodesics and the quasiglobal coordinate

In our further discussion, we will frequently use the quasiglobal coordinate ρ (see Section 3.1), which possesses some advantages as compared with other choices of the radial coordinate.

Thus, we can say in advance that this ρ behaves at horizons in the same way as the coordinates used to carry out an analytic extension of the metric, and therefore it can be used simultaneously on both sides of the horizon (hence the name "quasiglobal"). In addition, as we will see immediately, in some important cases the behavior of the ρ coordinate can be a criterion of geodesic completeness of the space-time and therefore the very necessity of its extension.

In terms of the quasiglobal coordinate $u = \rho$, the static metric (3.1) is written as follows:

$$ds^2 = A(\rho)dt^2 - \frac{d\rho^2}{A(\rho)} - r^2(\rho)d\Omega^2, \tag{3.30}$$

where $A(\rho) \equiv e^{2\gamma}$.

We have previously introduced the notion of a horizon using as examples the de Sitter, Schwarzschild and other metrics in which the coordinate $r = \rho$ turned out to be simultaneously a curvature and quasiglobal coordinate. In agreement with what has been said, let us define a horizon in any space-time with the metric (3.30) as a regular sphere $\rho = \rho_h$, near which

$$A(\rho) \sim (\rho - \rho_h)^n, \tag{3.31}$$

where $n \in \mathbb{N}$ is the order of the horizon. Since $A(\rho)$ is equal to the norm $\xi^\alpha \xi_\alpha$ of the Killing vector $\xi^\alpha = (1, 0, 0, 0)$, regular surfaces where $A = 0$ are Killing horizons (surfaces where the timelike or spacelike Killing vector becomes null).

It is easy to notice that, in terms of the metric (3.30), an R-region is a region where $A > 0$, and a T-region is a region where $A < 0$. A horizon of odd order n separates an R-region from a T-region, while a horizon of even order separates two R-regions or two T-regions from one another.[7]

Consider the behavior of geodesics of the metric (3.30) near horizons. According to (3.28), in the notation of the metric (3.30) ($u = \rho$, $e^{2\gamma} = e^{-2\alpha} = A(\rho)$), and bearing in mind that $\dot{u} = du/d\lambda$, we obtain

$$\pm\frac{d\lambda}{d\rho} = \left[E^2 - A\left(\frac{L^2}{r^2} + k\right)\right]^{-1/2}. \tag{3.32}$$

In particular, as $\rho \to \rho_h$ such that $A(\rho_h) = 0$ (a possible horizon), where the quantity ρ_h may be finite or infinite, we have

$$d\lambda \approx E\, d\rho. \tag{3.33}$$

If E is a finite constant, then, evidently, near a possible horizon the coordinate ρ behaves as a canonical parameter for any geodesic approaching it, be it spacelike, timelike or null. For timelike geodesics one can state that *a horizon is achieved along a geodesic at finite proper time $\lambda \equiv \tau$ if and only if this horizon corresponds to a finite value of the quasiglobal coordinate ρ.*

If $A \to 0$ as $\rho \to \infty$, such a surface may be called a *remote horizon*: it is reached along a geodesic at infinite proper time, and apparently any nongeodesic trajectories cannot make this time finite.

The constant E can be zero only in the case (see (3.24)) of spacelike geodesics completely located in the spatial section $t = \text{const}$. In this case the canonical parameter λ, coinciding with the proper length l, is determined by the integral

$$l = \int \frac{du}{\sqrt{A(1 - L^2/r^2)}}. \tag{3.34}$$

[7] Noninteger n would violate the analyticity of the metric even if the curvature invariants are finite, thus making meaningless an analytic extension beyond the horizon. Thus, if $\rho > \rho_h$ in a R- or T-region, then at fractional n the metric loses sense at $\rho < \rho_h$. Although while considering only a static reference frame one could include fractional n into consideration, we here put $n \in \mathbb{N}$ for simplicity.

The integrand is meaningless at $r < L$. Therefore $r = L$ is the minimum value of r reached by purely spatial geodesics. This is the simple geometric meaning of the constant L.

According to (3.32), if $E = 0$, then in the limit $A \to 0$, if $\rho \to \infty$, then the canonical parameter λ along the geodesics tends to infinity, just as for nonzero values of E. Thus we obtain the following important result for static, spherically symmetric space-time:

If $\rho \to \infty$ on a surface where $A \to 0$ (i.e., on a candidate horizon), then this surface (called a remote horizon) is the boundary of the space-time under consideration, and it cannot be reached by any geodesics at finite values of their canonical parameter.

So the space-time is geodesically complete at a remote horizon, and no extension is needed beyond such a horizon.

Let us now return to usual (not remote) horizons and look how the coordinate time t (it is also the time according to the clocks of a distant observer at rest in the case of an asymptotically flat space-time) behaves there on timelike and null geodesics approaching such a horizon. From (3.24) and (3.28) we obtain for the metric (3.30)

$$\frac{d\rho}{dt} = \pm A\sqrt{1 - V(\rho)/E^2}, \tag{3.35}$$

where $E > 0$, and the potential has the form $V = A(k + L^2/r^2)$. At the horizon, the potential turns to zero, and due to (3.31), we get from (3.35)

$$t \sim \pm \int \frac{d\rho}{(\rho - \rho_h)^n} \to \pm\infty. \tag{3.36}$$

Thus *the coordinate time t is infinite at the horizon for all geodesics that cross it, except for purely spatial ones*: it is equal to $+\infty$ for motion to the horizon and $-\infty$ for motion from the horizon. It means that a horizon is in absolute past or absolute future for any observer located in the static region, and it becomes clear that to cross a horizon and to gain a joint description of regions on its both sides it is necessary to pass on to coordinates which do not belong to the static RF. More than that: even from this consideration, completely restricted to the static RF, it follows that a horizon as a limiting surface in the set of spheres $r = \text{const}$ splits into two parts: *the past horizon* and *the future horizon*.

Eq. (3.35) looks especially simple for radially moving photons: $d\rho/dt = \pm A$. It shows that *the coordinate velocity of photons (and the same for massive particles), from the viewpoint of any static observer, tends to* *zero as the photon approaches a horizon*, and tends so rapidly that the coordinate time t at which a photon reaches or leaves a horizon is infinite.

We see that if a given static region of space-time contains a future horizon, then this horizon and the region beyond it realize the notion of a BH from the viewpoint of this static region since, to put it simply, it is possible to fly there but it is impossible in principle to return from there — from the absolute future. In a similar way, a past horizon and a region beyond it can be called a *white hole*, from which photons or any other particles can appear, but it is impossible in principle to get there from the static region because it is located in the absolute past.

3.3.3 Transitions to Lemaître reference frames

Let us now use the geodesic equations to make a transition to a nonstatic RF connected with a family of test particles moving in a static, spherically symmetric gravitational field.

Consider the general static metric (3.30) written using the quasiglobal coordinate ρ. We shall see that a transition from the coordinates (ρ, t) of the static RF to the geodesic coordinates (R, τ), where τ is the proper time along a geodesic and the radial coordinate R is the congruence parameter, different for different geodesics, can be described in a general form. A radial timelike geodesic in the metric (3.30) satisfies the equations following from (3.24) and (3.28)

$$\left(\frac{d\rho}{d\tau}\right)^2 = E^2 - A(\rho), \qquad \frac{dt}{d\tau} = \frac{E}{A(\rho)}, \tag{3.37}$$

where the constant E is connected with the initial velocity of a particle moving along this particular geodesic at a given value of R. In general, $E = E(R)$, i.e., the particle energies are different for different geodesics.

Equation (3.37) gives two of the four components of the transition matrix

$$\|\partial(t, \rho)/\partial(\tau, R)\|, \tag{3.38}$$

namely, $\dot\rho$ and $\dot t$ (dots and primes here stand for $\partial/\partial\tau$ and $\partial/\partial R$, respectively) since this partial differentiation occurs along the geodesics:

$$\dot\rho = \pm\sqrt{E^2(R) - A(\rho)}, \qquad \dot t = E(R)/A(\rho). \tag{3.39}$$

A relation between the other two components, t' and ρ', can be found from the condition $g_{\tau R} = 0$ when we substitute $dt = \dot t d\tau + t' dR$ and

$d\rho = \dot{\rho} d\tau + \rho' dR$ into the metric (3.30):

$$t' = \frac{\sqrt{E^2(R) - A(\rho)}}{E(R)A(\rho)} \rho'. \tag{3.40}$$

It remains to find $\rho'(R, \tau)$, which can be done by using the integrability condition $(\partial_\tau \partial_R - \partial_R \partial_\tau)\rho = 0$. The latter takes the form of a linear first-order differential equation with respect to ρ' considered as a function of R and ρ, $\rho' = y(R, \rho)$:

$$\partial_\rho y = -\frac{y\, \partial_\rho A}{2(E^2 - A)} + \frac{EE'}{E^2 - A}. \tag{3.41}$$

Solving it, we obtain

$$y = \rho'(R, \tau) = \sqrt{E^2 - A(\rho)} \left[f_0(R) + EE' \int \frac{d\rho}{\big(E^2 - A(\rho)\big)^{3/2}} \right], \tag{3.42}$$

where $f_0(R)$ is an R-dependent integration constant. The other integrability condition $(\partial_\tau \partial_R - \partial_R \partial_\tau)t = 0$ holds automatically due to (3.40). The functions $t(R, \tau)$ and $\rho(R, \tau)$ can now be found by further integration of Eqs. (3.39)–(3.42). The resulting metric can be written as follows:

$$ds^2 = d\tau^2 - \frac{\rho'(R, \tau)^2}{E^2(R)} dR^2 - r^2(\rho(R, \tau)) d\Omega^2. \tag{3.43}$$

One can see how this procedure works using the de Sitter space-time as an example. Its static form is (3.30) with $r \equiv \rho$ and $A(r) = 1 - H^2 r^2$, $H = \text{const} > 0$. Let us choose three different families of geodesics such that

$$E(R) = \sqrt{1 - kR^2}, \quad k = 0, \pm 1, \tag{3.44}$$

and show that their corresponding reference frames represent the three well-known forms of the de Sitter metric as isotropic cosmologies with different signs of spatial curvature (see the end of the previous chapter). Indeed, integrating the first relation in (3.39) as an equation for $\tau = \tau(r, R)$ and properly choosing the arbitrary function of R that appears as an integration constant, we obtain the following expressions for r:

$$r(R, \tau) = (R/H) \times \{\cosh(H\tau),\, \mathrm{e}^{H\tau}, \sinh(H\tau)\}, \tag{3.45}$$

where the expressions in the curly brackets are ordered according to $k = 1$, $0, -1$. Substituting them into (3.43), we obtain the metric in the form

$$ds^2 = d\tau^2 - a^2(t) \left(\frac{dR^2}{1 - kR^2} + R^2 d\Omega^2 \right),$$

$$a(\tau) = (1/H) \times \{\cosh(H\tau),\, \mathrm{e}^{H\tau}, \sinh(H\tau)\}, \tag{3.46}$$

as was intended. One can also verify that the expression (3.42) with $E(R)$ given by (3.44) (provided the function $f_0(R)$ is chosen properly) coincides with the expression for r' obtained directly from (3.45) in all three variants. So the transition in the metric has been completed.

We see that the de Sitter metric in the form of an isotropic cosmological model is written in geodesic RFs, comoving to some families of freely moving test particles in de Sitter space-time.

Thus for finding the metric in the Lemaître RF we needed only the transition matrix (3.38) consisting of derivatives whereas for writing explicitly the corresponding coordinate transformation it is necessary to perform further integration. In our example this integration has been performed for r as a function of R and τ but not for t. Let us find $t(R,\tau)$ for the geodesic family $k=0$ that leads to a spatially flat cosmology. We have the expressions

$$\dot{t} = \frac{1}{1 - R^2\,\mathrm{e}^{2H\tau}}, \qquad t' = \frac{R\,\mathrm{e}^{2H\tau}}{1 - R^2\,\mathrm{e}^{2H\tau}}. \tag{3.47}$$

Their integration gives

$$t = t_0 + \tau - \frac{1}{2H}\ln(1 - R^2\,\mathrm{e}^{2H\tau}), \tag{3.48}$$

where t_0 is an integration constant corresponding to an arbitrary zero point of t. The corresponding expression for r from (3.45) is $r = (R/H)\,\mathrm{e}^{H\tau}$.

The resulting metric (3.46) does not show any trace of the horizon at $r = H$ that restricted the static RF. This agrees perfectly with the previous result that a horizon is reached by a geodesically moving particle at its finite proper time. Such a particle (or observer) simply does not notice a horizon being crossed. On the other hand, according to (3.48), the coordinate time of the static RF makes sense only at $r < H$ and diverges where $r = H$. (This happens at different values of the proper time τ of different geodesics whose family parameter R characterized their initial conditions.)

Meanwhile, the metric (3.46) makes sense at any r. Thus the transformation to a geodesic (Lemaître) RF is one of the ways of extending the metric beyond a horizon, and such an extension was really carried out not only for de Sitter but also for BH space-times. It is really a useful tool for many purposes, in particular, for comparison and matching vacuum metrics with those describing the gravitational field inside matter distributions (such as, e.g., the Tolman solution for a dust cloud, to be presented in Sec. 3.6.2). It is, however, not very convenient for extracting information on the global structure of space-time.

3.3.4 Horizons, R- and T-regions

The horizons described above in static (or homogeneous if a T-region is concerned) spherically symmetric space-times are special cases of more general kinds of horizons discussed in the literature on the theory of gravity. Let us mention some of them, only briefly explaining their meaning but without going into deep details and subtleties since it would require an introduction of many concepts from such advanced branches of mathematics as differential and algebraic topology and global differential geometry. An interested reader can find a detailed analysis of these concepts and their application to gravity in the books [168, 198, 296, 413] and in review articles.

An *event horizon* is a boundary of a black hole, and both these notions are rigorously defined in an asymptotically flat space-time, while in other cases, such as spaces with a nonzero cosmological constant, natural extensions of these notions are used. It is known (Penrose's theorem) that an event horizon is formed by null geodesics.

An *apparent horizon* is defined as a two-dimensional surface, such that null geodesics which orthogonally leave it have zero expansion. In stationary BHs, the apparent horizon coincides with the event horizon; however, an advantage of the apparent horizon is that it is defined locally.

A *Cauchy horizon* is the boundary of a space-time region in which the evolution of physical fields and matter distributions can be found from their equations of motion with initial data specified on a certain Cauchy hypersurface (which can be null or spacelike). The emergence of Cauchy horizons is most frequently connected with singularities since they can be a source of impacts unpredictable in the framework of classical gravity.

A *Killing horizon* is, by definition, a surface whose normal is null and coincides with a Killing vector. In other words, it is a surface at which a certain Killing vector is null. Quite evidently, Killing horizons can occur only in space-times possessing isometries which are then characterized by Killing vectors.

For static space-time, of greatest importance are Killing horizons at which the timelike Killing vector $\xi^\mu = (1, 0, 0, 0)$ becomes null. This happens where $g_{00} = e^{2\gamma} = 0$. From the viewpoint of a static RF, there is a singularity at such a horizon, although the 4-curvature is certainly finite[8].

[8] The definition of a Killing horizon implicitly assumes its regularity since otherwise such a definition would be meaningless: we would then deal with a singularity rather than a horizon.

Lastly, sometimes the notion of a horizon is introduced, connected with a certain scalar. Thus, for instance, in spherically symmetric space-times with the metric (3.1), the quantity $r(u,t)$ is a scalar from the viewpoint of the two-dimensional subspace parametrized by the coordinates u and t. On the basis of the behavior of the function $r(u,t)$, one can introduce the notions of R- and T-regions, generalizing these notions from static to arbitrary nonstatic space-times. By definition, an **R-region** is a space-time region where the gradient of the function $e^{\beta} = r(u,t)$, considered as a scalar in the (u,t)-subspace, is spacelike:

$$r^{,\alpha} r_{,\alpha} = e^{-2\gamma}\dot{r}^2 - e^{-2\alpha} r'^2 < 0. \tag{3.49}$$

In this case the quantity r can be chosen as a spatial coordinate. A **T-region** is, by definition, a region where $r^{,\alpha} r_{,\alpha} > 0$, and therefore one can choose r as a time coordinate. Furthermore, an r**-horizon** is a surface on which $r^{,\alpha} r_{,\alpha} = 0$.

All the horizons considered previously are evidently r-horizons and Killing horizons. The sphere $r = 2m = 2GM$ in the Schwarzschild metric and the sphere $r = r_+$ in the Reissner–Nordström metric are Killing horizons and event horizons simultaneously, while $r = r_-$ in the Reissner–Nordström metric is a Cauchy horizon, like many horizons in metrics with $\Lambda \neq 0$.

3.4 Schwarzschild black holes. Geodesics and a global description

3.4.1 R- and T-regions

As already noticed, in the Reissner–Nordström–de Sitter solution (3.17) the radial coordinate is simultaneously a curvature coordinate and a quasiglobal coordinate ($r \equiv \rho$). Therefore we can directly use the results of the previous section in analyzing the properties of the solution.

Let us address the simplest special case, the Schwarzschild metric (3.20). At $r = 2GM = 2m = r_g$ (this value of r is called the gravitational radius corresponding to the mass M, or the Schwarzschild radius), there is a simple horizon, and in the region $r < 2GM$ (T-region) the coordinates r and t exchange their roles: r becomes a temporal coordinate and t a spatial one (just as was already discussed for the de Sitter metric).

The T-region is described by the Kantowski–Sachs metric

$$ds^2 = \frac{dr^2}{|A(r)|} - |A(r)|dy^2 - r^2 d\Omega^2, \qquad (3.50)$$

where $|A(r)| = 2GM/r - 1$, and the spatial coordinate y replaces t. If we suppose that r decreases from the value $2GM$ to zero, then the homogeneous model (3.50) contracts (collapses) to zero in two angular directions along the coordinate spheres, whereas in the third direction (y) it stretches without limit. The final state $r = 0$ is a spacelike singularity, reached at a certain time instant.[9]

The inverse process of anisotropic evolution (expansion of coordinate spheres $\theta = \text{const}$, $\phi = \text{const}$ and contraction from infinity to zero along the y axis) is observed as r changes from $r = 0$ to $r = 2GM$.

The metric (3.50) describes both expanding (in the sense of growing r) and contracting cosmological models since the time direction is not unambiguously related to the growth or fall of r: the physical time τ is related to r by the integral

$$\tau = \pm \int dr/\sqrt{A(r)}, \qquad (3.51)$$

whose convergence at both $r \to 2GM$ and $r \to 0$ indicates that the evolution from the horizon to the singularity (or conversely) occurs in a finite time interval according to a clock at rest in the metric (3.50) at a fixed spatial point (y, θ, ϕ).

Meanwhile, the expanding or contracting T-region is in no way connected with the initial R-region even though the curvature at the horizon $r = \tau = 2GM$ which separates them is finite: the Kretschmann scalar is

$$\mathcal{K} = 48(GM)^2/r^6. \qquad (3.52)$$

3.4.2 Geodesics in the R-region

The behavior of geodesics near the horizon corresponds to the general description presented above. Thus, according to (3.28) and (3.24), in the

[9] It is sometimes claimed that the Schwarzschild space-time has a singularity at the centre. As is clear from the above-said, this claim is completely wrong. Indeed, the singularity $r = 0$ is in a T-region, which represents a homogeneous cosmological model, whose all spatial points are equivalent, whereas the notion of a centre assumes a certain distinguished spatial point. Moreover, in the coordinate system in which the metric (3.50) is written, all spatial points reach the singularity simultaneously.

Schwarzschild metric

$$\left(\frac{dr}{d\lambda}\right)^2 + \left(1 - \frac{2m}{r}\right)\left(\frac{L^2}{r^2} + k\right) = E^2,$$

$$\frac{dt}{d\lambda} = \frac{E}{1 - 2m/r}, \quad m := GM. \tag{3.53}$$

(The quantity m, having the dimension of length, is sometimes called the geometrized mass.) The first of these equations gives near the horizon $(r \to 2m)$, independently of the nature of the geodesic and on the angular momentum L, $dr/d\lambda \approx \pm E$, whence (assuming $E > 0$) $d\lambda \approx \pm dr/E$. Thus *all geodesics cross the horizon at a finite value of their canonical parameter* λ. In particular, for timelike geodesics it means that *the horizon is crossed at finite proper time* s.

In the exceptional case of purely spatial geodesics ($E = 0$), (3.34) implies:

$L > 2m$: $r = L$ is the minimum value of r, such geodesics do not reach the horizon.

$L = 2m$: the horizon is reached but the length $l \sim \int dr/(r - 2m)$ logarithmically diverges. In other words, the distance from any fixed point to the horizon along such a (tangent) geodesic is infinite.

$L < 2m$: the horizon is reached, and the integral (3.34) converges at $r = 2m$.

It is also not difficult to verify directly that the 3-curvature of the hypersurfaces $t = \text{const}$ is singular at the horizon.

Let us dwell upon massive particle motion in the Schwarzschild field, i.e., on timelike geodesics ($k = 1$, $d\lambda = ds$), where ds is the increment of the particle's proper time. From (3.53) we obtain

$$\left(\frac{dr}{ds}\right)^2 + V(r) = E^2 \tag{3.54}$$

with the potential

$$V(r) \equiv \left(1 + \frac{L^2}{r^2}\right)\left(1 - \frac{2m}{r}\right). \tag{3.55}$$

Taking the derivative d/ds, we obtain the equation of motion

$$\frac{d^2 r}{ds^2} = -\frac{1}{2}V'(r). \tag{3.56}$$

Stable circular orbits can be found from the conditions $V' = 0,\ V'' > 0$. Substituting the expression of the potential, we obtain

$$mr^2 - L^2 r + 3mL^2 = 0. \tag{3.57}$$

A solution of this quadratic equation gives extrema of the potential:

$$r_{\text{ext}}(L) = \frac{L}{2m}(1 \pm \sqrt{1 - 12m^2/L^2}). \tag{3.58}$$

The sign of the second-order derivative $V''(r)$ at $r = r_{\text{ext}}$

$$V''(r_{\text{ext}}) = \frac{2}{r^4}(2mr - L^2)\Big|_{r=r_{\text{ext}}}$$

distinguishes the orbital radii ($V'' > 0$ [the plus sign in (3.58)] and a maximum of the potential that occurs at smaller radii [the minus sign in (3.58)]:

$$r = r_{\max}(L) = \frac{L}{2m}(1 - \sqrt{1 - 12\,m^2/L^2}) = \frac{6m}{1 + \sqrt{1 - 12\,m^2/L^2}}. \tag{3.59}$$

It is clear that stationary orbits are absent at small angular momenta L,

$$L \leq L_{\min} = 2\sqrt{3}\,m.$$

The minimum orbital radius is

$$r_{\min} = r_{\text{ext}}(L_{\min}) = 6GM = 3r_g, \tag{3.60}$$

and this orbit is only marginally stable due to $V'' = 0$ (see Fig. 3.1).

Let us stress that in curved space-time the radius r of a coordinate sphere is in general not equal to its distance from the centre even if there is a centre. And in the Schwarzschild field, as we have seen, there is no centre at all because the value $r = 0$ belongs to a T-region.

3.4.3 Particle capture by a black hole

Let us return to Eq. (3.54) and consider the problem of particle capture by a Schwarzschild black hole. Let the particle move from infinity with an "energy" $E \geq 1$ (where $E = 1$ corresponds to zero velocity at infinity). If it is not captured by the black hole, it again escapes to infinity. Then there should be a maximum proximity position such that

$$dr = 0 \Rightarrow V(r) = E^2.$$

On the contrary, if it is captured, the equation $V(r) = E^2$, with $V(r)$ given by (3.55), must not have a solution.

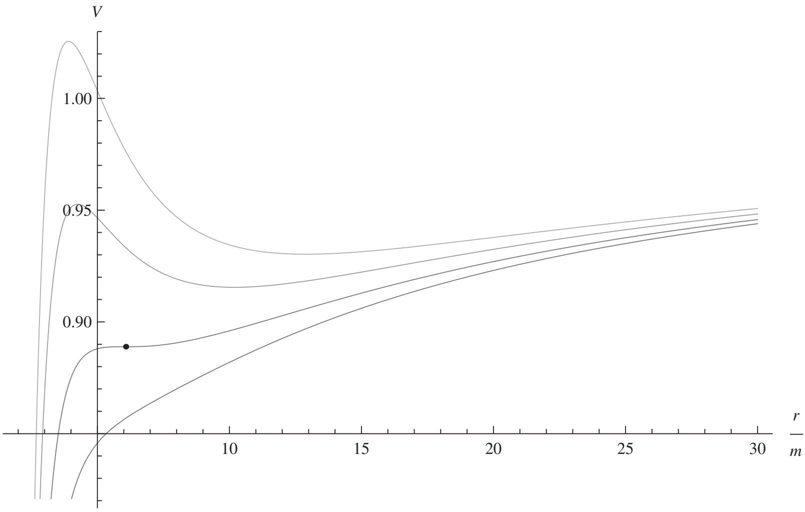


Figure 3.1 The potential (3.55) for $L/m = 3.2, 2\sqrt{3}, 3.8, 4.1$ (bottom-up). Stable circular orbits correspond to minima of the potential. The dot on the second curve at $r/m = 6$ corresponds to the smallest (marginally stable) circular orbit. The horizon is located at $r/m = 2$.

For a particle of unit mass moving from infinity with the initial velocity $v_\infty \ll 1$, we have the following expressions for the energy and angular momentum corresponding to nonrelativistic mechanics in flat space: $E_\infty = 1 + v_\infty^2/2 \simeq 1$ and $L = bv_\infty$, where b is the impact parameter. We arrive at the algebraic equation

$$\left(1 + \frac{L^2}{r^2}\right)\left(1 - \frac{r_g}{r}\right) \simeq 1,$$

whose solution is absent if

$$L \lesssim 2r_g \Rightarrow b \lesssim 2r_g/v_\infty.$$

Consequently, the capture cross-section for a nonrelativistic particle is

$$\sigma = \pi b^2 = 4\pi r_g^2/v_\infty^2 = 16\pi m^2/v_\infty^2. \tag{3.61}$$

In the general, relativistic case, the capture condition is the same: the absence of solutions to Eq. (3.54) with $dr/ds = 0$. By definition of the angular momentum (see (3.26)),

$$L \equiv r^2 \frac{d\phi}{ds}.$$

At large distances from the black hole, $\phi \approx b/r$, hence $d\phi = -dr(b/r^2)$ (where again b is the impact parameter), and therefore $dr/ds = L/b$. Substituting it into Eq. (3.54), at large r, we obtain

$$E^2 = 1 + \frac{L^2}{b^2}.$$

Now, for arbitrary r, (3.54) takes the form

$$\left(1 + \frac{L^2}{r^2}\right)\left(1 - \frac{r_g}{r}\right) = 1 + \frac{L^2}{b^2}.$$

In the ultrarelativistic limit, $L/r \gg 1$, $L/b \gg 1$, and the equation is simplified:

$$\frac{b^2}{r^2}\left(1 - \frac{r_g}{r}\right) \simeq 1.$$

This equation has no solution if

$$b \leq \frac{3\sqrt{3}}{2} r_g.$$

Thus the capture cross-section for ultrarelativistic particles moving near a Schwarzschild black hole is

$$\sigma = \pi b^2 = \frac{27}{4}\pi r_g^2. \tag{3.62}$$

The above derivations did not take into account that an interaction of a particle with a black hole is accompanied by gravitational wave emission. Most frequently this emission is really negligible, but it cannot be neglected at interactions of two black holes and, in such cases, it substantially increases the cross-section of their mutual capture. More precisely, the cross-section of gravitational capture of two black holes with masses M_0 and M to a close binary is expressed as follows [304]:

$$\sigma = 2\pi \left(\frac{85\pi}{6\sqrt{2}}\right)^{2/7} \frac{G^2 (M_0 + M)^{10/7} M_0^{2/7} M^{2/7}}{c^{10/7} v_{\rm rel}^{18/7}}, \tag{3.63}$$

where $v_{\rm rel}$ is the relative velocity of the two black holes. After the capture, the black holes rapidly merge due to energy loss from gravitational wave emission, and a new mighty gravitational-wave outburst is produced at the moment of merging. The above cross-section is larger than the cross-section of direct black hole collisions taking into account the gravitational focusing in a wide range of black hole masses and velocities.

3.4.4 A global description: The Kruskal metric

A transition from the Schwarzschild static metric to coordinates providing its complete description by a single map was independently formulated in 1960 by Kruskal and Szekeres [261, 394].

Let us describe this transition. Only the (r, t) subspace should be transformed while the angular coordinates θ, ϕ remain the same.

First of all we transform r passing on to the "tortoise" coordinate r_*, such that the two-dimensional metric in the (r, t) subspace becomes conformally flat:

$$
\begin{aligned}
r_* &= \int \left(1 - \frac{2m}{r}\right)^{-1} dr = r + 2m \ln \left|\frac{r}{2m} - 1\right|, \\
ds_2^2 &= \left(1 - \frac{2m}{r}\right) dt^2 - \left(1 - \frac{2m}{r}\right)^{-1} dr^2 \\
&= \left(1 - \frac{2m}{r}\right)(dt^2 - dr_*{}^2).
\end{aligned}
\tag{3.64}
$$

The horizon corresponds to the limit $r_* \to -\infty$.

In Eq. (3.64) it is convenient to pass on to the null coordinates V, W:

$$
2t = V + W, \quad 2r_* = V - W. \tag{3.65}
$$

At fixed V, the horizon corresponds to $W \to \infty$, hence $t \to \infty$. On the contrary, at fixed W the horizon corresponds to $V \to -\infty$, hence $t \to -\infty$. Thus (in agreement with the previous discussion of geodesics), the horizon is either in absolute past or in absolute future with respect to an observer at rest and splits into a past horizon and a future horizon. So, in terms of V and W, the Schwarzschild metric has the form

$$
ds^2 = \left(1 - \frac{2m}{r}\right) dV\, dW - r^2 d\Omega^2. \tag{3.66}
$$

We have not yet removed the singularity at $r = 2m$. To this end, one more step is necessary: a transformation of the form $V = V(v)$, $W = W(w)$. Let us for certainty consider a transition through the future horizon: we fix V and concentrate on the limit $W \to \infty$. We wish to choose such a function $W(w)$ that the coefficient of $dV dw$ in the metric is finite. It is easy to verify that a suitable choice is $W = -2m \ln |w|$, because $1 - 2m/r = (r - 2m)/r \sim r_*/(2m) \sim -w/(2m)$ as $r \to 2m$.

In a similar way the metric is regularized at a transition through the past horizon. It is not hard to see that the transformation

$$V = 2m \ln|v|, \quad W = -2m \ln|w| \tag{3.67}$$

leads to the relation

$$1 - \frac{2m}{r} = -vw\, \mathrm{e}^{-r/(2m)}, \tag{3.68}$$

and the metric acquires the regular form

$$ds^2 = \frac{32m^3}{r}\, \mathrm{e}^{-r/(2m)} dv\, dw - r^2 d\Omega^2. \tag{3.69}$$

It is possible to further transform the null coordinates into spatial (R) and temporal (T) ones by assuming

$$v = \frac{1}{2}(T+R), \quad w = \frac{1}{2}(T-R), \quad dv\, dw = dT^2 - dR^2, \tag{3.70}$$

whence

$$ds^2 = \frac{32m^3}{r}\, \mathrm{e}^{-r/(2m)} (dT^2 - dR^2) - r^2 d\Omega^2. \tag{3.71}$$

This is the Kruskal metric giving a complete picture of the Schwarzschild space-time. The initial region $r > 2m$ is mapped in it into the right quadrant $v > 0, w < 0$, or $R > 0$, $-R < T < R$ (region R_+, see Fig. 3.2). In addition to it, there are two T-regions T_- and T_+ (the first one is

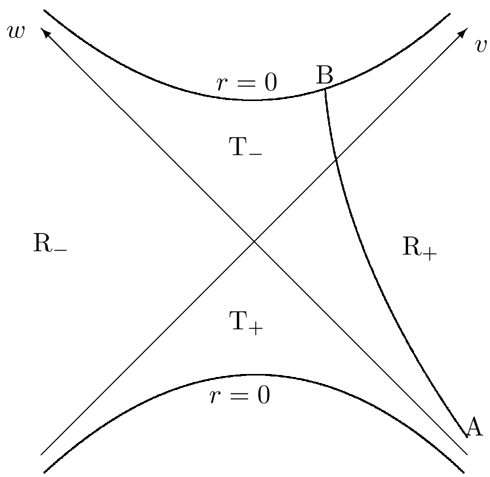


Figure 3.2 The Kruskal diagram.

obtained by transition through the future horizon, the second one through the past horizon) and one more R-region, R_-.

The horizon $r = 2m$ is now depicted by the pair of intersecting axes $v = 0$ and $w = 0$, or, equivalently, $T = \pm R$. These axes (or these null hypersurfaces if we recall that each point on the diagram corresponds to a 2-sphere) do not belong to any of the R- or T-regions, and the same is true for their intersection point $u = v = 0$; still all points at the horizons are regular points in the Schwarzschild-Kruskal space-time, as is evident from the metric (3.69) or (3.71).

The singularity $r = 0$ is depicted by two hyperbolas: $T = -\sqrt{1 + R^2}$ (the past singularity) and $T = +\sqrt{1 + R^2}$ (the future singularity). It is the future singularity where everything gets after falling beyond the future event horizon.

Due to the null nature of the v and w coordinates, all radial null geodesics in the diagram are parallel to the v or w axes. Other causal curves (timelike as well as nonradial null trajectories) have slopes smaller than 45° from the vertical. This is true, in particular, for the lines $r =$ const in the R-regions (which depict the coordinate spheres at rest and are not geodesics).

3.4.5 From Kruskal to Carter-Penrose diagram for the Schwarzschild metric

In studies of the causal structure of space-time containing Killing horizons, it is often helpful to use Kruskal-like coordinate transformations which provide smooth transitions between regions separated by horizons. The results are most frequently presented with the aid of two-dimensional diagrams (Carter–Penrose diagrams) whose form contains the basic information on the global causal structure of the space-time in question. Such diagrams are more convenient than the Kruskal diagram because each R- or T-region is depicted there by a square or triangle of fixed finite size.

Let us build such a diagram for the Schwarzschild space-time using the Kruskal metric (3.69):

$$ds^2 = \frac{32m^3}{r}\, \mathrm{e}^{-r/(2m)} dv\, dw - r^2 d\Omega^2, \tag{3.72}$$

where the radius r is related to the coordinates v and w by Eqs. (3.68), or

$$-vw = \frac{r - 2m}{2m}\, \mathrm{e}^{r/(2m)}. \tag{3.73}$$

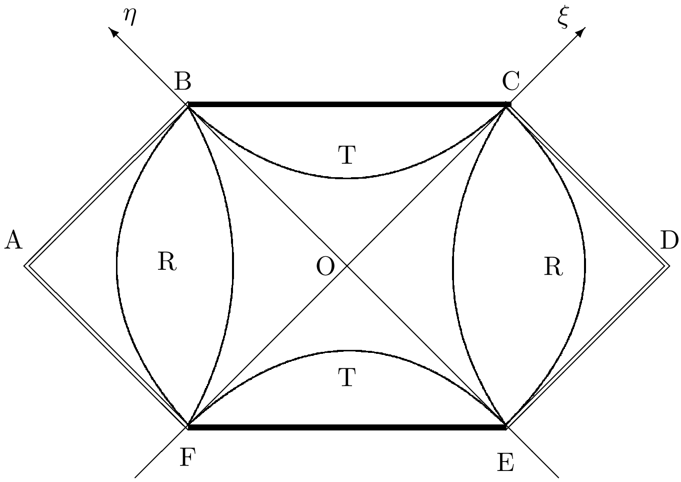


Figure 3.3 The Carter–Penrose diagram for the Schwarzschild metric.

The diagram becomes compact if, instead of the coordinates $v \in \mathbb{R}$ and $w \in R$, we introduce the coordinates ξ, η, specified on the segment $(\pi/2, \pi/2)$ by putting

$$v = \tan\xi, \quad w = \tan\eta. \tag{3.74}$$

It is easily seen (Fig. 3.3) that the horizon is now represented by the two segments $\xi = 0$, $\eta = 0$ (BE and CF), and the singularity $r = 0$ by other two segments where $\xi + \eta = \pm\pi/2$ (BC and EF). The values $\xi = \pm\pi/2$ and $\eta = \pm\pi/2$ (the line strings FAB and CDE) correspond to an infinite radius r. The whole two-dimensional (r, t) manifold has been mapped into the interior of the hexagon ABCDEF. The null nature of the coordinates ξ and η allows for easily distinguishing spacelike, timelike and null directions at any point in precisely the same way as it is done using the Kruskal diagram.

3.5 The global causal structure of space-times with horizons

3.5.1 Crossing the horizon in the general case

Transformations similar to those applied to the Schwarzschild metric are of quite general nature, and, as a result, Carter–Penrose diagrams can be built in the general case of two-dimensional sections of space-time with the aid of simple correspondence rules between the behavior of metric functions and graphic images.

Consider an arbitrary static or stationary space-time where the metric depends on a spatial coordinate u, and let us restrict ourselves to its ***two-dimensional section*** with the metric

$$ds_2^2 = \mathrm{e}^{2\gamma(u)}dt^2 - \mathrm{e}^{2\alpha(u)}du^2. \qquad (3.75)$$

(If the metric depends on other spatial coordinates, we suppose that they are fixed.) Let the surface $u = u^*$ be a horizon, so that $\mathrm{e}^{\gamma(u^*)} = 0$. Let us introduce two "radial" coordinates x and ρ by the relations

$$\begin{aligned} dx &= \pm\,\mathrm{e}^{\alpha-\gamma}du \\ &\Rightarrow ds_2^2 = \mathrm{e}^{2\gamma}(dt^2 - dx^2), \quad x = \text{“tortoise”}. \end{aligned} \qquad (3.76)$$

$$\begin{aligned} d\rho &= \pm\,\mathrm{e}^{\alpha+\gamma}du, \\ ds_2^2 &= \mathrm{e}^{2\gamma}dt^2 - \mathrm{e}^{-2\gamma}d\rho^2, \quad \rho = \text{quasiglobal}. \end{aligned} \qquad (3.77)$$

As $u \to u^*$, at the horizon $x \to \pm\infty$ [it is easy to verify that a finite $x(u^*)$ would mean a singularity, see (3.7)], while the coordinate ρ can tend both to a finite and to an infinite limit. As we have already seen in Sec. 3.3, if $\rho \to \pm\infty$ as $u \to u^*$, the space-time under consideration is geodesically complete at $u \to u^*$ even without an extension beyond the horizon, and such a "remote horizon" is a natural boundary of the space-time. No extension is necessary in this case.

Let us address the more general case $|\rho(u^*)| < \infty$. Without losing generality we suppose that $\rho(u^*) = 0$, $x \to -\infty$ and $\rho \to +0$ as $u \to u^*$. Besides, we assume that in some finite neighborhood of $u = u^*$

$$\mathrm{e}^{2\gamma} = \rho^q\, F(\rho), \qquad (3.78)$$

where $q = \text{const} \geq 1$, and $F(\rho)$ is an analytic function with a finite value $F(0)$. It is precisely this behavior of the metric that takes place in most situations of physical interest.

By analogy with the method of obtaining the Kruskal picture, we first introduce the null coordinates

$$V = t + x, \quad W = t - x. \qquad (3.79)$$

The limit $V \to -\infty$ at fixed finite W corresponds to the past horizon (since $t \to -\infty$) while the limit $W \to \infty$ at fixed finite V to a future horizon. Further, if one introduces new null coordinates v and w related to V and W by

$$V = V(v), \quad W = W(w), \qquad (3.80)$$

then, under some simple requirements on these functions, the mixed coordinate patch (v, W) will cover the past horizon $V \to -\infty$ while the (V, w) patch will cover the future horizon $W = +\infty$.

Indeed, consider the future horizon. It is easy to verify that a finite value of the metric coefficient g_{Vw} at $w = 0$ is achieved if, as $W \to \infty$,

$$w \sim \begin{cases} \mathrm{e}^{-WF(0)/2}, & (q = 1), \\ W^{-q/(q-1)}, & (q > 1). \end{cases} \tag{3.81}$$

A transition across the future horizon corresponds to a smooth transition of the w coordinate through zero.

On the other hand, it is also easy to verify that, under the assumptions made, as $x \to -\infty$, the following relationship between ρ and x is valid:

$$\rho \sim \begin{cases} \mathrm{e}^{xF(0)/2} & (q = 1), \\ |x|^{-1/(q-1)} & (q > 1). \end{cases} \tag{3.82}$$

Consequently, according to (3.78), at transitions through horizons *the coordinate ρ behaves in precisely the same way as w*, i.e., smoothly crosses the zero value. Thus precisely this choice of the static spatial coordinate makes it possible to describe the region beyond the horizon, despite the impossibility of describing the whole space-time in the framework of the static reference frame. It is for this reason that the coordinate ρ is called quasiglobal.

Depending on the parity of q, the region beyond the horizon will be a R- or T-region.

The extension through the past horizon is carried out in absolutely the same way with the aid of the (v, W) patch and leads to similar results.

Remark. It often happens in specific problems that it is hard to perform a transition to the quasiglobal coordinate in the whole region. For instance, it can be necessary to solve a transcendental equation. Let us note in this connection that for the above reasoning it is only significant that the coordinate ρ satisfy the relation (3.77) *near the horizon*, remaining arbitrary in the remaining space. This makes the choice much easier.

We have seen that an extension of the metric across a horizon can be described in terms of a certain distinguished coordinate (w or ρ), which changes its sign at such a transition. If, however, the number q is fractional in the representation (3.78), then at the values $\rho < 0$ the metric coefficients lose their meaning. Hence the candidate horizon is actually a space-time singularity. Such a singularity is connected with *a loss of analyticity of the* *metric*, and it does not necessarily imply an infinite value of any algebraic invariant of the curvature tensor. In such cases, evidently, some *differential* invariants, containing derivatives of the form $(\nabla_\alpha \nabla_\beta \ldots) R_{\mu\nu\rho\sigma}$, are singular.

We will sometimes call such surfaces *singular horizons*.

This "quantization" condition $q \in \mathbb{N}$ can turn out to be not unique if fractional powers of ρ appear in other metric coefficients besides $g_{\rho\rho}$ and g_{tt}. Such examples are really sometimes considered in the literature.

3.5.2 Construction of Carter–Penrose diagrams

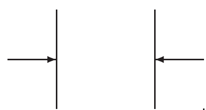
As follows from the above, for an analysis of the global causal structure of two-dimensional Lorentzian surfaces with the metric (3.75), it is convenient to use the coordinate ρ, specified by (3.77) and connected with the null coordinates providing a smooth extension of the metric. Thus, suppose that under some choice of the coordinates, the metric of a two-dimensional section of space-time has the form

$$ds^2 = f(\rho)dt^2 - \frac{d\rho^2}{f(\rho)}, \tag{3.83}$$

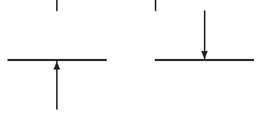
where the coordinate ρ belongs to some finite or infinite range $a < \rho < b$ on the real axis. According to (3.76), $x = \pm \int d\rho / f(\rho)$. Then, using the same reasoning as was used above in the discussion of the global properties of the Schwarzschild space-time, it is not difficult to verify that in some coordinate system one can juxtapose the end points a and b and zeros ($\rho = \rho_i$) of the function $f(\rho)$ to the following graphic images (assuming that there are zeros of $f(\rho)$ not coinciding with a or b):

$\rho \to a, \quad \rho \to b$:

$f > 0, \ |x| < \infty$

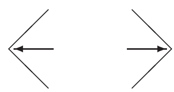


$f < 0, \ |x| < \infty$

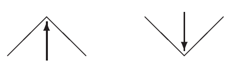


$\rho \to a, \ \rho \to b, \ \rho \to \rho_i$:

$f \to +0, \ |x| \to \infty$



$f \to -0, \ |x| \to \infty$



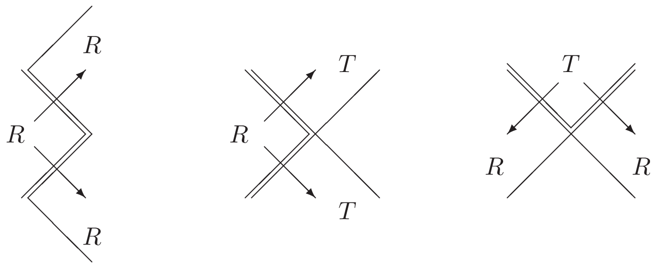


Figure 3.4 Examples of transition across horizons. The double tilted segments show the horizon of the initial region, single segments show new segments added at the transition, and arrows correspond to the direction of the transition. The symbol R designates a region where $f(\rho) > 0$ (an R-region), the symbol T a region where $f(\rho) < 0$ (a T-region). In a few subsequent figures the notations are the same.

(the arrows mark the direction along which we approach the point a, b or ρ_i). The construction process begins from any of the end points. To each interval where the function $f(\rho)$ has a constant sign there corresponds the interior of a triangle or a square in the diagram. A horizon $\rho = \rho_i$ is depicted by a pair of tilted segments forming a right angle. To carry out a transition across this horizon, one should add two more tilted segments of the same length as shown in Fig. 3.4. The way of adding new segments is chosen depending on the sign of $f(\rho)$ at $\rho < \rho_i$ and $\rho > \rho_i$. The new pair of segments corresponds to the same value $\rho = \rho_i$.

One can formulate a convenient rule for transitions through horizons: **at RT and TR transitions** (i.e., from a R-region to a T-region or vice versa) **the old and new segments form an X-shaped figure while at RR and TT transitions the figure is like a zigzag.** An explanation of the rule is that a horizon with a given value of ρ_i is located in the diagram to the right or to the left of a R-region (and splits into a past horizon and a future horizon) whereas with respect to a T-region a horizon is located either above or below it.

The construction can continue endlessly or terminate if the diagram covers a finite area whose boundary points correspond to the boundary values $\rho = a$ and $\rho = b$, which (recall) can be finite or infinite.

For each pair of regions separated by a tilted segment (i.e., a horizon) there exists a common system of null coordinates like (3.79). The rules formulated above are sufficient for building any two-dimensional Carter–Penrose diagrams as soon as the metric is written in the form (3.83). To illustrate the method, let us give examples of diagrams for the

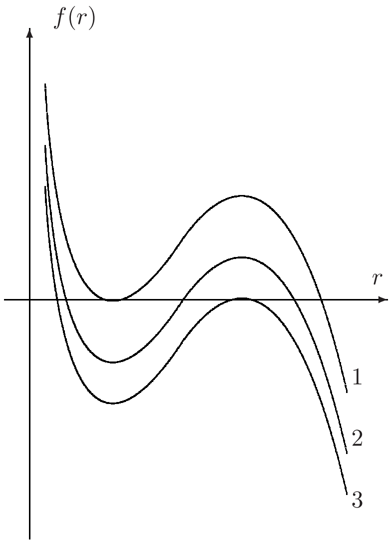


Figure 3.5 The qualitative behavior of the function $f(r)$ in the metric (3.84) at $0 < \Lambda q^2 < 1/4$ with two or three zeros.

Reissner–Nordström–dS metric, for which, in the notations (3.83),

$$f(r) = 1 - 2m/r + q^2/r^2 - \Lambda r^2/3, \tag{3.84}$$

where m and q are the mass and the charge, respectively, in geometrized units. The coordinate ρ introduced in (3.77) coincides here with the (Schwarzschild) radial coordinate r.

The behavior of the function $f(r)$ at $r \in (0, \infty)$ is rather diverse and depends on the values of the parameters m, q, Λ. Figure 3.5 shows the qualitative form of $f(r)$ in the cases where it has three simple roots or one simple root and one double root. Figures 3.6–3.8 show the corresponding Carter–Penrose diagrams. In other, simpler cases where $f(r)$ has a single root or two simple roots, the space-time structure is well known, see, e.g., the books [198, 296]. With the aid of the rules given above, the reader can easily build the corresponding diagrams him or herself.

In Figs. 3.6–3.8, the tilted segments depict the Killing horizons whose corresponding roots r_i of the function $f(r)$ are enumerated in increasing order. The squares excluded from the diagram in Fig. 3.7 as well as the segments drawn by thick lines in Figs. 3.6 and 3.8 can play the role of branching points (a trip around any of then does not necessarily lead to the

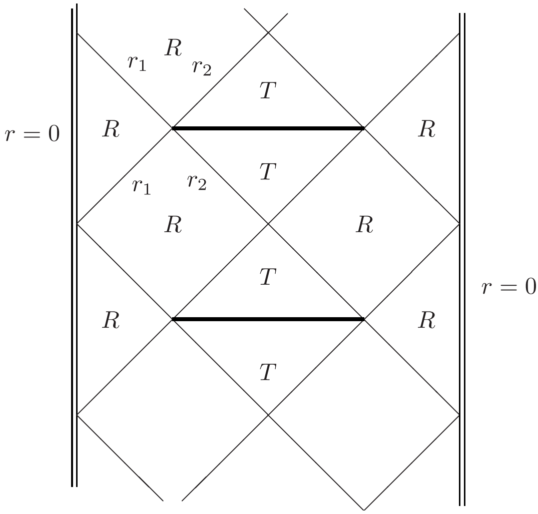


Figure 3.6 The Carter–Penrose diagram corresponding to curve 1 for $f(r)$ in Fig. 3.5. The diagram continues indefinitely both up and down. The thick lines correspond to $r = \infty$.

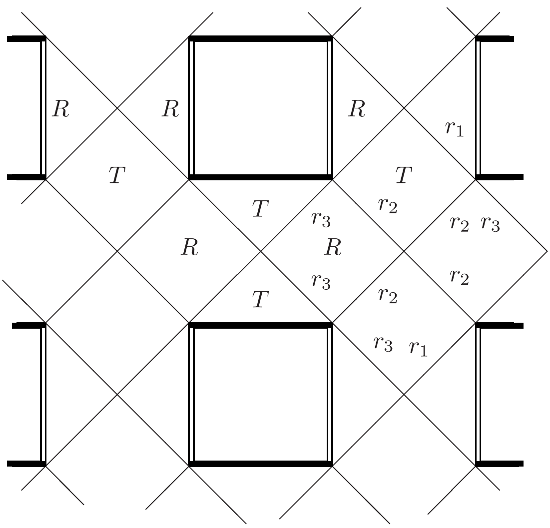


Figure 3.7 The Carter–Penrose diagram corresponding to curve 2 for $f(r)$ in Fig. 3.5. The diagram continues indefinitely in all directions and occupies the whole plane except the squares bounded by double and thick lines. The double lines correspond to $r = 0$, the thick ones to $r = \infty$.

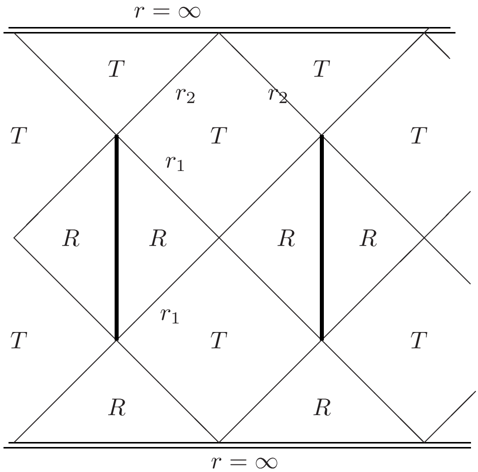


Figure 3.8 The Carter–Penrose diagram corresponding to curve 3 for $f(r)$ in Fig. 3.5. The diagram continues indefinitely to the left and to the right. The thick lines correspond to $r = 0$.

same initial region), therefore any multisheet extensions of these diagrams are possible. On the other hand, identifications of isometric hypersurfaces are also possible.

3.6 A black hole as a result of gravitational collapse

3.6.1 Internal and external regions. Birkhoff's theorem

As is known from astrophysics, as soon as the nuclear fuel of a star is exhausted, it experiences a catastrophic contraction (collapse) due to gravitational forces which can no longer be balanced by hydrodynamic and radiative pressure. The result of a gravitational collapse depends on which mass is taking part in it (generally speaking, it is not the star as a whole but only its inner part). At masses smaller than the Chandrasekhar limit (about 1.4 $M_\odot$, where $M_\odot$ is the solar mass), the stellar inner core becomes a white dwarf with a density of the order of 10^6 $\text{g}\cdot\text{cm}^{-3}$. At $m > 1.4 m_\odot$, after gravitational collapse there emerges a neutron star whose density is of the order of nuclear matter density, (10^{13}–10^{15} $\text{g}\cdot\text{cm}^{-3}$), while at still larger masses the mass of the collapsing core exceeds the stability limit of neutron stars (about 3–4 $M_\odot$, its precise value depends on the equation of state of neutron matter, on the structure details of the neutron star and on the choice of the theory of gravity), and the contraction leads to BH emergence. As shown by calculations, a BH should form if the initial stellar mass is $\gtrsim 10M_\odot$; in this process, more than half of the stellar mass is thrown into the ambient space, which appears to a remote observer as a supernova explosion.

BHs of smaller mass could form from collapsing density fluctuations at very early stages of the Universe evolution.

Calculations of the gravitational collapse are very complicated, there are only a few exactly solvable models (such as, for instance, the collapse of a spherically symmetric dust cloud [398]; see, e.g., [263]), while in more complex cases the problem is solved only numerically. Still there are some general, model-independent features of gravitational collapse ending with BH formation.

The main feature is that outside the collapsing body the space-time remains a vacuum, hence all the results obtainable for the properties of vacuum solutions to the equations of gravity are valid. Moreover, if a spherically symmetric body is collapsing, the whole process occurs with spherical symmetry, and then Birkhoff's famous theorem holds [37],[10] whose simplified formulation is "in GR, a spherically symmetric gravitational field in vacuum is necessarily static". It then reduces to the Schwarzschild solution (or Schwarzschild–de Sitter if the cosmological constant is included).[11] From a more general and correct viewpoint, the theorem indicates the case where the field equations, under certain conditions, induce an additional space-time symmetry that was not postulated initially. For example, for a T-region of the same Schwarzschild space-time, what follows from the theorem is certainly not staticity (which is absent) but homogeneity, or independence of the metric on the spatial coordinate which corresponds to the temporal coordinate t in the static region.

[10] After Birkhoff the theorem was extended to spherically symmetric systems with a cosmological constant Λ, systems with electromagnetic and scalar fields, some scalar-tensor theories (STT) of gravity, etc. The most general approach to the problem [78] consisted in studying the general conditions under which one could prove the static or homogeneous nature of the system. This included into consideration not only all the previously studied cases of the theorem in GR and STT but also many new ones. The theorem was generalized in two respects: inclusion of new symmetry types (planar, cylindrical, pseudoplanar) and new kinds of matter sources of gravity (scalar fields, gauge fields, perfect fluid, etc.). A multidimensional generalization of this approach is described in [80].

[11] The theorem was actually proved earlier by J. Jebsen [217], but it has happened that his work is less well-known than Birkhoff's; see also [218].

Let us present a schematic proof of Birkhoff's theorem for the simplest case of spherical symmetry [the metric (3.1)] and vacuum, $T_\mu^\nu \equiv 0$. According to the Einstein equations, $R_\mu^\nu \equiv 0$. Suppose that the inequality (3.49) holds in the space-time region under consideration, that is, we are in a R-region, and we can choose $r = e^\beta$ as a spatial coordinate. Then in the expressions (3.3) for the Ricci tensor components one can put $\dot\beta = 0$. The equation $R_{01} = -2\dot\alpha\beta' = 0$ then immediately leads to $\dot\alpha = 0$ [$\beta' \neq 0$ due to (3.49)], and among the three metric functions only γ can still be time-dependent. The Einstein equation $G_1^1 = R_1^1 - \frac{1}{2}R = 0$ contains γ only in the combination $e^{-2\alpha}\beta'\gamma'$, whence it follows that $\gamma' \equiv d\gamma/dr = \gamma'(r)$, and, $\gamma = \gamma_1(r) + \gamma_2(t)$. But the function $\gamma_2(t)$ can be excluded by a proper choice of the time coordinate, which leads to a completely static metric. The theorem is proved.[12]

The validity of Birkhoff's theorem is related to the fact that in GR the gravitational field has spin 2 and consequently it does not have dynamic monopole and dipole degrees of freedom. In other words, there are no monopole and dipole gravitational waves.

As follows from Birkhoff's theorem, in the collapse of spherical bodies, the exterior space-time is unique and is described by the Schwarzschild metric at all times. From the properties of the metric it follows that, from the viewpoint of a distant observer, the stellar surface contracts to the horizon for an infinitely long time — it seems to be frozen at the Schwarzschild radius $r = 2m$. However, as follows from calculations, deflections from $r = 2m$ decay exponentially in the distant observer's time, with a characteristic time of about $2m$, which is about 10μs (the time for which light passes $2m = 3$ km) in the case of $m = M_\odot$ [340]. Very quickly it happens that signals from the collapsing stellar surface stop coming, and it disappears from the sight of a remote observer almost instantaneously.

On the surface itself, only a finite proper time passes at contraction from any finite radius, not only to the horizon but to the singularity $r = 0$ (recall that it is a cosmological singularity).

[12]If, in the conditions of the theorem, we replace the inequality (3.49) with its opposite, i.e., assume that we are in a T-region, then, by the same reasoning, we shall again find out that the metric is t-independent, but t is now a spatial coordinate, and the independence from it proves the space-time homogeneity. However, if one assumes that the gradient of r is null, the theorem cannot be proved, and under this condition wavelike solutions to the Einstein equations become possible.

The global space-time structure for a collapsing star drastically differs from the structure of an "eternal" Schwarzschild BH depicted in Kruskal (Fig. 3.2) and Carter–Penrose (Fig. 3.3) diagrams. Thus, for instance, in Fig. 3.2, from the whole empty space-time there only remains the region to the right of the curve AB that depicts the world line of a point on the stellar surface, while everything to the left of this curve remains uncertain until a dynamic solution to the Einstein equations is known for the stellar interior. There remain only one R-region (the right one) and one T-region (the upper, contracting one).

3.6.2 Gravitational collapse of a spherical dust cloud

As an example of a solution describing the interior of a collapsing body, let us consider Tolman's solution [398] describing the evolution of a spherical cloud of dustlike matter in a comoving reference frame. Since for dust particles this is also a geodesic reference frame, the metric can be taken in the synchronous form

$$ds^2 = d\tau^2 - \,\mathrm{e}^{2\lambda(R,\tau)} dR^2 - r^2(R,\tau) d\Omega^2, \tag{3.85}$$

where τ is the proper time along the particle trajectories labelled by different values of the radial coordinate R. The only nonzero component of the SET is $T_0^0 = \rho$, and the Einstein equations read

$$2r\ddot{r} + \dot{r}^2 + 1 - e^{-2\lambda} = 0, \tag{3.86}$$

$$\frac{1}{r^2}(1\dot{r}^2 + 2r\dot{r}\dot{\lambda}) - \frac{\mathrm{e}^{-2\lambda}}{r^2}(2rr'' + r'^2 - 2rr'\lambda') = 8\pi G\rho, \tag{3.87}$$

$$\dot{r}' - \dot{\lambda} r' = 0. \tag{3.88}$$

Equation (3.88) is easily integrated in τ giving

$$\mathrm{e}^{2\lambda} = \frac{r'^2}{1 + f(R)}, \tag{3.89}$$

where $f(R)$ is an arbitrary function satisfying the condition $1 + f(R) > 0$. Substituting (3.89) into (3.86), we obtain the equation

$$2r\ddot{r} + \dot{r}^2 = f(R), \tag{3.90}$$

whose first integral is

$$\dot{r}^2 = f(R) + \frac{F(R)}{r}. \tag{3.91}$$

This expression makes clear the physical meaning of the function $f(R)$: since $\dot{r}$ can be understood as the radial velocity of a dust particle, $f(R)$ specifies the initial dust velocity distribution: for $f \geq 0$ it is the particle velocity squared at large r ($f > 0$ and $f = 0$ correspond to hyperbolic and parabolic motion, respectively). If $f(R) < 0$ but $F(R) > 0$ (elliptic motion), the particle cannot reach infinity, and the quantity $F(R)/|f(R)|$ shows the maximum value of r accessible to it.

The meaning of the other arbitrary function, $F(R)$, becomes clear if we substitute (3.89) and (3.91) into (3.87). We obtain

$$\rho = \frac{1}{8\pi G}\frac{F'(R)}{r^2 r'}, \tag{3.92}$$

or

$$F(R) = 8\pi G \int \rho r^2 r' dR. \tag{3.93}$$

Assuming that in the initial configuration (before collapse) there was a regular centre, so that $F = 0$ at $r = 0$, we can write

$$F(R) = 2GM(R), \quad M(R) = 4\pi \int_0^r \rho r^2 dr, \tag{3.94}$$

so that $M(R)$ is the mass function equal to the mass of a spherical body including all matter inside the sphere of given radius r, and $F(R)$ is the corresponding Schwarzschild radius.

Eq. (3.91) can be further integrated, and the solution can be written in a parametric form for $f \neq 0$ and explicitly for $f = 0$ [263]:

$$f > 0: \quad r = \frac{F}{2f}(\sinh\eta - 1), \quad \tau_0 - \tau = \frac{F}{2f^{3/2}}(\sinh\eta - \eta); \tag{3.95}$$

$$f < 0: \quad r = \frac{F}{-2f}(1 - \cos\eta), \quad \tau_0 - \tau = \frac{F}{2(-f)^{3/2}}(\eta - \sin\eta); \tag{3.96}$$

$$f = 0: \quad r = \left(\frac{9F}{4}\right)^{1/3}(\tau_0 - \tau)^{2/3}. \tag{3.97}$$

The new arbitrary function $\tau_0(R)$ here indicates the instant of τ at which the particle with a given R reaches the singularity $r = 0$. To find out whether this singularity occurs in a R-region (and can then be called a singular centre) or in a T-region (and is then cosmological in nature), we must check whether the gradient of $r(R, \tau)$ is spacelike or timelike at small r. Evidently,

$$r^{,\alpha}r_{,\alpha} = \dot{r}^2 - \mathrm{e}^{-2\lambda}r'^2 = -1 - 2r\ddot{r}, \tag{3.98}$$

where the second equality sign follows from (3.86).

It can be verified that for any $f(R)$ near the singularity the function $r(R,\tau)$ is given by Eq. (3.97), whence it follows that $r\ddot{r} \sim -(\tau_0 - \tau)^{-2/3} \to -\infty$, so that $r^{,\alpha}r_{,\alpha} \to +\infty$, and the singularity occurs in a T-region for any choice of the arbitrary functions.

Then of interest is the question: when does a given dust particle cross the horizon thus getting from R- to T-region? The answer again follows from (3.98): since $\ddot{r} = F(R)/(2r^2)$, the gradient of r is null where $r = F(R) = 2GM(R)$, that is, a particle crosses the horizon precisely at the time instant when $r(R,\tau)$ reaches the value of the Schwarzschild gravitational radius corresponding to the interior of the sphere specified by a given R. Note that such a crossing is quite regular, and this instant is actually unnoticed in the comoving reference frame.

To obtain a global solution including both a dust distribution and an external vacuum region, it is necessary to describe such a vacuum (Schwarzschild) region in a reference frame suitable for smooth matching to the internal solution. This problem is easily solved since the vacuum solution in a proper form is obtained from the Tolman solution by putting $F(R) = \mathrm{const}$ whence $\rho = 0$. The resulting solution reproduces the Schwarzschild metric in a geodesic reference frame as described above in Sec. 3.3.3.

So a spherical collapse of a dust cloud always leads to a BH. One of the most fundamental questions discussed in BH physics is whether or not there are matter configurations whose gravitational collapse can lead to a naked singularity instead of a BH. The so-called cosmic censorship conjecture requires that such unwanted entities as space-time singularities, if any, can only appear separated from external observers by BH horizons. There are counterexamples to censorship, processes where naked singularities do occur as a result of collapse, but it is still not certain whether or not such situations are generic. Moreover, an answer can be different in GR and its different extensions; see, in addition to textbooks on gravity and BH physics, the reviews [124, 187, 209, 216, 414].

This page intentionally left blank

Chapter 4

Black holes under more general conditions

4.1 Black holes and massless scalar fields

In the previous sections we have seen that the presence of a massless vector (electric or magnetic) field does not prevent the existence of BHs. In other words, a BH can possess an electric or/and magnetic charge and be a source of the corresponding fields. In this section we will discuss the question of whether a BH can also be a source of a massless scalar field. To this end, it appears to be possible to use exact solutions which are in this case obtained rather easily and whose properties are of separate interest.

4.1.1 The general STT and the Wagoner transformations

In what follows we will consider the properties of static, spherically symmetric gravitational fields for two most important versions of the massless scalar field equations, the minimally coupled one[13] ($\psi = \phi^{\rm min}$) with the equation $\Box\psi = 0$ and the conformally coupled one ($\phi = \phi^{\rm conf}$) with the equation $\Box\phi + \frac{1}{6}R\phi = 0$, where R is the scalar curvature.

[13]The term "minimally coupled" applied to any field means (see Chapter 2) that the influence of gravity manifests itself in replacing partial derivatives in the field equations by covariant ones.

We would like to begin with a more general formulation of nonminimal coupling of scalar fields to gravity called the general (Bergmann–Wagoner–Nordtvedt) scalar-tensor theory (STT) of gravity, whose Lagrangian

$$L_{\rm STT} = f(\phi)\tilde{R} + h(\phi)\tilde{g}^{\mu\nu}\phi_{,\mu}\phi_{,\nu} - 2U(\phi) + L_m, \tag{4.1}$$

is specified in a space-time $\tilde{V}_4$ with the metric $\tilde{g}_{\mu\nu}$ and scalar curvature $\tilde{R}$. $U(\phi)$ is a potential, in some cases generalizing a cosmological constant; L_m is the Lagrangian of all other matter which can also depend on $\tilde{g}_{\mu\nu}$ and ϕ.

Let us present the well-known conformal mapping [412] which formally reduces rather a wide class of systems with scalar fields in the framework of the Bergmann–Wagoner–Nordtvedt STT of gravity to systems with the scalar field $\phi^{\min}$. Consider the Lagrangian of the general STT (4.1) specified in the space-time $\tilde{V}_4$ with the metric $\tilde{g}_{\mu\nu}$ and the scalar curvature $\tilde{R}$. L_m is the Lagrangian of the rest of matter, maybe also depending on ϕ and $\tilde{g}_{\mu\nu}$. This formulation, corresponding to the ***Jordan conformal frame*** (or ***picture***) is of fundamental nature in STT as it is usually formulated since it is in this picture that the SET of matter T^μ_ν satisfies the usual conservation law $\nabla_\nu T^\nu_\mu = 0$, thus providing the standard equations of motion in the atomic system of measurement. In particular, free particles move along geodesics.

The field equations are easier studied and solved in the so-called ***Einstein conformal frame*** (or ***picture***) in which the scalar field is minimally coupled to gravity. Namely, the conformal transformation [412]

$$\tilde{g}_{\mu\nu} = \frac{g_{\mu\nu}}{f(\phi)}, \quad \frac{d\phi}{d\psi} = f\left|fh + \frac{3}{2}\left(\frac{df}{d\phi}\right)^2\right|^{-1/2}, \tag{4.2}$$

leads the Lagrangian (4.1) to a Lagrangian typical of GR (up to a full divergence) with $\psi = \phi^{\min}$:

$$L = R + \varepsilon g^{\mu\nu}\psi^{,\mu}\psi_{,\nu} + \frac{-2U(\phi) + L_m}{f^2(\phi(\psi))}, \quad \varepsilon := \operatorname{sign}\left[fh + \frac{3}{2}\left(\frac{df}{d\phi}\right)^2\right] \tag{4.3}$$

for a system with the scalar field ψ in a space-time with the metric $g_{\mu\nu}$. It is supposed that the function $f(\phi)$ used in the transformation (4.2) is finite and positive in the whole region of interest.

Depending on $\varepsilon = \pm 1$, the STT splits into two classes: normal ones, in which $\varepsilon = +1$, i.e., the kinetic energy of the scalar field in the Einstein picture is positive, and anomalous or phantom (ghost), $\varepsilon = -1$, where this energy is negative.

Special examples of the theory (4.1) are GR with a minimally coupled scalar field ($f \equiv 1,\ h = \varepsilon = \pm 1$) and a conformally coupled scalar field $\phi = \phi^{\rm conf}$:

$$f = 1 - \phi^2/6, \qquad h = 1, \qquad \varepsilon = +1, \quad \sqrt{6}\phi = \tanh\frac{\psi+\psi_0}{\sqrt{6}}, \tag{4.4}$$

where ψ_0 is an arbitrary constant.

The equation for a massless conformally coupled scalar field

$$\Box\phi + \frac{1}{6}R\phi = 0 \tag{4.5}$$

is invariant under the conformal transformation $\tilde g_{\mu\nu} = \Omega^{-2}(x) g_{\mu\nu}$, accompanied by the scalar field transformation $\tilde\phi = \Omega\phi$, as one can verify directly using the corresponding transformation formulas for the scalar curvature and the d'Alembert operator:

$$\begin{aligned} \tilde R &= \Omega^2 R - 6\Omega\,\Box\,\Omega + 12\Omega^{,\alpha}\Omega_{,\alpha}, \\ \tilde\Box\phi &= \Omega^2\,\Box\,\phi - 2\Omega\Omega^{,\alpha}\phi_{,\alpha}. \end{aligned} \tag{4.6}$$

It is easy to see that the transformation (4.2) is different from the one with respect to which Eq. (4.5) is invariant. Therefore it becomes possible to pass on from Eq. (4.5) to the equation $\Box\psi = 0$.

The most well-known example of an STT is the Brans-Dicke theory [49] for which

$$f(\phi) = \phi, \qquad h(\phi) = \frac{\omega}{\phi}, \quad \phi = \exp[(\psi-\psi_0)/\sqrt{|\omega+3/2|}], \tag{4.7}$$

$\omega = \text{const}$. This theory is normal for $\omega > -3/2$, phantom for $\omega < -3/2$ and has no dynamics at $\omega = -3/2$.

On phantom fields

We have seen that a large class of STT reduce to a phantom field in GR. Being quite unusual, such fields are often considered as purely artificial entities. But recently these fields attracted much interest in connection with the Dark Energy problem, a search for models able to account for the so-called Dark Energy (DE) that drives the observed growing acceleration of the Universe expansion (for more details see Part II of this book, in particular, the footnote on page 217).

It is known that phantom DE behavior can be explained without explicitly introducing phantom fields, in particular, in the framework of the general STT (4.1) which has sufficient freedom for describing all observational data (see, e.g., [169]).

The main objection against a phantom field is that a field with a negative kinetic term can have an arbitrarily large negative energy of high-frequency oscillations, which, according to quantum field theory, can lead to runaway production of particles and antiparticles accompanied by addition of equal negative energy of the phantom field itself. Nothing of this kind is observed, which casts serious doubt on possible existence of phantom fields.

On the other hand, there are theoretical reasons for considering phantom fields since they naturally appear in some models of string theory [367], supergravities [307] and theories in more than 11 dimensions like F-theory [236] and in some Kaluza-Klein theories (e.g., those discussed in Part III of this book). They can be effective fields, following in some reduction procedure from an underlying theory; they may not be subject to quantization (at least directly) and possibly do not interact with other fields. This, at least partly, counters the above objections.

Even not being a steady theoretical object, phantom fields seem useful for a kind of *gedanken experiment*, they can lead to models or predictions of interest which can afterwards be obtained by other, less exotic means. It is the main reason for considering them in this book.

4.1.2 Minimally coupled scalar fields

Consider configurations with the metric (3.1) in the presence of a massless scalar field $\psi = \phi^{\rm min}$) with the Lagrangian

$$L_s = \frac{1}{2}\varepsilon\psi^{,\alpha}\psi_{,\alpha}, \quad \varepsilon = \pm 1, \tag{4.8}$$

and an electromagnetic field $F_{\mu\nu}$ with the Lagrangian

$$L_{\rm em} = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu}. \tag{4.9}$$

The most general form of the electromagnetic potential compatible with spherical symmetry is

$$A_\mu = \delta^0_\mu A(u) + \delta^3_\mu q_m \cos\theta + \partial\Phi/\partial x^\mu, \tag{4.10}$$

where $\Phi(x^\mu)$ is an arbitrary function and $q_m = \text{const}$ is the magnetic charge. Accordingly, the SETs of the fields ψ and $F_{\mu\nu}$ have the form

$$T^\nu_\mu[\psi] = \frac{1}{2}\varepsilon\, {\rm e}^{-2\alpha}\psi'^2\, {\rm diag}(1,-1,1,1); \tag{4.11}$$

$$T^\nu_\mu[F] = 2(\mathbf{E}^2 + \mathbf{B}^2)\, {\rm diag}(1,1,-1,-1), \tag{4.12}$$

where

$$\mathbf{E}^2 = F^{01}F_{10} = q_e^2/r^4; \quad \mathbf{B}^2 = F^{23}F_{23} = q_m^2/r^4, \tag{4.13}$$

$\mathbf{E}$ are $\mathbf{B}$ are the electric field strength and the magnetic induction, and

$$F^{01} = q_e\, \mathrm{e}^{-\alpha-\gamma}/r^2, \quad F_{10} = q_e\, \mathrm{e}^{\alpha+\gamma}/r^2, \tag{4.14}$$

where q_e is the electric charge. Let us denote $Q^2 = \frac{1}{2}\varkappa(q_e^2 + q_m^2)$.

The equation for ψ and its first integral are

$$(\,\mathrm{e}^{\gamma+2\beta-\alpha}\psi')' = 0, \quad \psi' = \bar{C}\,\mathrm{e}^{\alpha-2\beta-\gamma}, \tag{4.15}$$

where the integration constant $\bar{C}$ is interpreted as a scalar charge.

The tensors (4.11) and (4.12) have the property $T_1^1 + T_2^2 = 0$. In this case it is helpful to use the harmonic ($\Box u = 0$) radial coordinate u [50], such that the metric coefficients in (3.1) satisfy the condition (see Sec. 3.1)

$$\alpha = 2\beta + \gamma. \tag{4.16}$$

Then the combination $\binom{1}{1} + \binom{2}{2}$ of the Einstein equations reduces to the Liouville equation

$$(\beta + \gamma)'' = \mathrm{e}^{2\beta+2\gamma}, \tag{4.17}$$

which is easily integrated giving

$$\mathrm{e}^{-\beta-\gamma} = s(k,u) := \begin{cases} k^{-1}\sinh ku, & k > 0; \\ u, & k = 0; \\ k^{-1}\sin ku, & k < 0; \end{cases} \tag{4.18}$$

where k is an integration constant; one more integration constant has been removed by choosing the zero point of u. As a result, we come to the metric

$$ds^2 = \mathrm{e}^{2\gamma}dt^2 - \frac{\mathrm{e}^{-2\gamma}}{s^2(k,u)}\left[\frac{du^2}{s^2(k,u)} + d\Omega^2\right]. \tag{4.19}$$

Without loss of generality we assume $u \geq 0$, so that $u = 0$ corresponds to spatial infinity at which the metric is asymptotically flat, and we can put in addition $\gamma(0) = 0$ under a proper choice of scale along the time axis t.

At the asymptotic ($u \to 0$), the coordinate u behaves as $1/r$, so that

$$r = \mathrm{e}^{\beta} \approx 1/u; \quad \mathrm{e}^{\alpha} \approx 1/u^2; \quad \gamma' \to \gamma'(0) = -GM, \tag{4.20}$$

where M is the Schwarzschild mass, as follows from comparing the limit of large r (4.20) with the Schwarzschild metric.

It should be noted that many exact static, spherically symmetric solutions obtained by now refer to systems with $T_1^1+T_2^2=0$ (or can be reduced to them) and can therefore be found using the method described here.

The field equation for ψ in the metric (3.1) with the coordinate (4.16) gives

$$\psi = \bar{C}u + \psi_0, \quad \psi_0 = \text{const.} \tag{4.21}$$

Let us pass on, in addition to Q, to geometrized mass and scalar charge: $m = GM$ and $C = \sqrt{\varkappa}\bar{C}$ having the dimension of length.

Among the remaining Einstein equations only one is independent (corresponding to one remaining unknown function $\gamma(u)$), but it makes sense to write down two equations: the component $\binom{0}{0}$ from (3.6), having the Liouville form, and the component $\binom{1}{1}$ from (3.5), which does not contain second-order derivatives and is a first integral of other equations:

$$e^{2\alpha} R_0^0 = -\gamma'' = -Q^2 e^{-2\gamma}, \tag{4.22}$$

$$e^{2\alpha} G_1^1 = \beta'(\beta' + 2\gamma') = \frac{1}{2}\varepsilon\psi'^2 - Q^2 e^{2\gamma}. \tag{4.23}$$

Integrating (4.22), we arrive at the following metrics:

(a) for $Q \neq 0$:

$$ds^2 = \frac{dt^2}{Q^2 s^2(h, u+u_1)} - \frac{Q^2 s^2(h, u+u_1)}{s^2(k,u)}\left[\frac{du^2}{s^2(k,u)} + d\Omega^2\right]; \tag{4.24}$$

with the constants k, h, u_1, C, Q connected with each other and the mass m by the relations

$$\begin{aligned} &s^2(h, u_1) = 1/Q^2; \\ &m^2 - Q^2 = h^2 \operatorname{sign} h = k^2 \operatorname{sign} k - \varepsilon C^2/2, \end{aligned} \tag{4.25}$$

where the first equality follows from the boundary condition $\gamma(0) = 0$ and the second one from substituting the expressions (4.18) for $\beta - \gamma$ and the solution to (4.22) into (4.23);

(b) for $Q = 0$:

$$ds^2 = e^{-2hu} dt^2 - \frac{e^{2hu}}{s^2(k,u)} \left[\frac{du^2}{s^2(k,u)} + d\Omega^2 \right]; \qquad (4.26)$$
$$Gm = h; \quad k^2 \operatorname{sign} k = \varepsilon C^2/2 + h^2.$$

The function s is defined by Eq. (4.18) in all cases.

In the limit $C \to 0$, the solution (4.21)–(4.25) passes on to the Reissner–Nordström electrovacuum solution and as $Q \to 0$ to the solution (4.26); in turn, (4.26) passes on, as $C \to 0$, to the Schwarzschild vacuum solution. In other coordinates the solution (4.26) (naturally, only for the normal sign of scalar field energy, $\varepsilon = 1$) was found for the first time by I.Z. Fisher in 1948 [164] and was repeatedly rediscovered later; the solution (4.21)–(4.25) was found by R. Penney in 1969 [324]. For $\varepsilon = -1$, the solution (4.26) was obtained in [35], and (4.21)–(4.25) in [50].

Let us briefly discuss the properties of solutions for $\varepsilon = +1$.

In the electrically neutral case (4.26) the coordinate u is changing in the range $u > 0$, and $u \to \infty$ corresponds to a naked singularity, which is attractive ($g_{00} \to 0$) for $m > 0$. The metric (4.24) behaves in a similar way at $h \geq 0$, i.e., at small charges as compared to the mass in proper units, $Q^2 \leq m^2$. The singularity may be characterized as scalar field dominated because there $\phi \to \infty$, and the total scalar field energy E_s is infinite whereas the electromagnetic field energy E_e is finite. The corresponding Reissner–Nordström electrovacuum metric possesses horizons (two for $Q^2 < m^2$ and one for $Q^2 = m^2$), which, as we see, disappear if there emerges a scalar field, however small the scalar charge C is.

At $h < 0$ (a large charge, $Q^2 > m^2$), the coordinate u is changing in the range $0 < u < u_{\max}$, where $u_{\max}$ is the closest zero of the function $\sin[|h|(u + u_1)]$. The value $u = u_{\max}$ corresponds to the central naked repulsive ($e^\gamma \to \infty$) singularity of Reissner–Nordström type, dominated by the electromagnetic field ($\phi < \infty$, $E_s < \infty$, $E_e = \infty$).

We see that in all cases the solution with a scalar field contains a naked singularity. Thus at least in spherically symmetric space-times the massless field $\phi^{\min}$ cannot be an external field of a BH. This statement, as we will see below, is one of the special cases of the so-called no-hair theorems for BH. Such a terminology traces back to J.A. Wheeler whose famous phrase “Black holes have no hair” meant that BHs have no external parameters other than mass, electric (and/or magnetic) charge and angular momentum.

4.1.3 Conformally coupled scalar field

The scalar-electrovacuum gravitational field with $\phi^{\rm conf}$ is obtained from (4.21)–(4.26) with the aid of the transformation (4.2), (4.4); as a result, the metric acquires a conformal factor substantially changing the nature of the geometry. The solution has the form:[14]

(a) for $Q \neq 0$

$$ds^2 = \frac{\cosh^2[C(u+u_0)]}{\cosh^2 Cu_0}\left\{\frac{dt^2}{q^2 s^2(h, u+u_1)} - \frac{q^2 s^2(h, u+u_1)}{s^2(k,u)}\left[\frac{du^2}{s^2(k,u)} + d\Omega^2\right]\right\}, \tag{4.27}$$

$$s^2(h, u_1) = 1/q^2, \quad q = Q\cosh Cu_0, \qquad k^2 \operatorname{sign} k = h^2 \operatorname{sign} h + 3C^2; \tag{4.28}$$

the electric field and the (geometrized) Schwarzschild mass m are determined according to (4.13) and by comparison with the Schwarzschild metric at large spherical radii $r = \sqrt{-g_{22}}$:

$$\mathbf{E}^2 = F^{10}F_{01} = q^2\,\mathrm{e}^{-4\beta} = q^2(g_{22})^{-2};$$

$$[m + C\tanh(Cu_0)]^2 = q^2 + h^2 \operatorname{sign} h; \tag{4.29}$$

(b) for $Q = 0$

$$ds^2 = \frac{\cosh^2[C(u+u_0)]}{\cosh^2 Cu_0}\left[\mathrm{e}^{-2hu}dt^2 - \frac{k^2\,\mathrm{e}^{2ku}}{\sinh^2 ku}\left(\frac{k^2 du^2}{\sinh^2 ku} + d\Omega^2\right)\right];$$

$$m = h - C\tanh Cu_0; \quad k = \sqrt{3C^2 + h^2}, \tag{4.30}$$

in both cases, the scalar field $\phi = \phi^{\rm conf}$ is given by

$$\phi = \sqrt{6}\tanh[C(u+u_0)]. \tag{4.31}$$

Let us note that, unlike the solution for $\phi^{\rm min}$, the coordinate condition (4.16) no longer holds because of the applied conformal transformation.

As before, the value of the coordinate $u = 0$ corresponds to flat spatial infinity.

In the case $h < 0$, the solution with a nonzero charge q (4.27)–(4.29) is specified on the segment $0 < u < u_{\max} < \infty$, where $u_{\max}$ is the zero of the

[14] We here restrict ourselves to normal scalar fields, $\varepsilon = +1$.

function $\sin[|h|(u+u_1)]$ closest to $u=0$. The value $u=u_{\max}$ corresponds to a naked central ($r=0$) repulsive ($g_{00}\to\infty$) singularity similar to the Reissner–Nordström one at $q^2>m^2$; however, unlike the analogous case with the field $\phi^{\min}$, here both energies E_e and $E_s^{\rm conf}$ diverge. The solution (4.27)–(4.29) at $h\geq 0$, $u_1<0$ behaves in a similar way.

The case $0<h<|C|$, $u_1>0$ differs from the one just described in that the solution now ranges over the whole positive semiaxis, $u>0$, and the singularity at $u=\infty$ is scalar-dominated: $r\to 0$, $e^{2\gamma}=g_{00}\to\infty$, $E_s^{\rm conf}=\infty$, $E_e<\infty$. The solution (4.30), (4.31) at $h<|C|$ behaves in a similar way.

In the case $|h|>|C|$, $u_1>0$, both solutions exist on the semiaxis $u>0$, but now the value $u=\infty$ corresponds to an attractive ($e^{2\gamma}\to 0$) naked singularity with an infinite spherical radius r, called a "space pocket" (the term is suggested by P. Jordan). The singular nature of this sphere is verified directly using (16). The spatial infinity ($u=0$) and the singularity ($u=\infty$) are separated by a throat, a sphere where the function $e^{\beta}=r(u)$ has a minimum. The energy $E_e<\infty$ (it is zero at $q=0$), whereas $E_s^{\rm conf}=-\infty$. The anomalous sign of the energy is related to the lack of positive definiteness of the temporal component of the SET for $\phi^{\rm conf}$.

At $h=|C|$, $u_1>0$, the sphere $u=\infty$ is regular, and it becomes necessary to extend the solution beyond it. We can recall that this regular sphere in the Jordan picture corresponds to a singularity in Einstein's picture. In other words, the whole space-time of Einstein's picture maps to only a part of the space-time in Jordan's picture. It is an example of the so-called conformal continuation [54].

To extend the solution beyond the sphere $u=\infty$, let us introduce the new radial coordinate

$$y=\coth|Cu|. \tag{4.32}$$

Then both solutions for $\phi^{\rm conf}$ take the form

$$\begin{aligned}
E^2 &= \sqrt{F^{10}F_{01}}=\frac{|q|}{r^2}\,\frac{|q|y^2}{h^2(y+y_0)^2(y+y_1)^2},\\
\phi &= \sqrt{6}\,\frac{1+yy_0}{y+y_0};\\
ds^2 &= (y+y_0)^2\left[\frac{dt^2}{(y+y_1)^2}-h^2\frac{(y+y_1)^2}{y^4}(dy^2+y^2d\Omega^2)\right];\\
y_0 &= \tanh hu_0;\qquad y_1=\coth hu_1.
\end{aligned} \tag{4.33}$$

At $Q = 0$ we have $y_1 = 1$; spatial infinity corresponds to $y = \infty$, where $\phi = \phi_\infty = \sqrt{6} y_0$.

If $y_0 < 0$, the solution is defined for $-y_0 < y < \infty$ and has a naked attracting singularity at the centre ($y \to -y_0 > 0$; $e^\gamma = e^\beta = 0$), where both fields ϕ and $F_{\mu\nu}$ are singular.

In the case $y_0 > 0$, the coordinate y is specified for all $y > 0$, so that, as $y \to 0$, just as in the limit $y \to \infty$, the metric becomes flat, i.e., there is the second spatial infinity where $e^{2\gamma} \to (y_0/y_1)^2$ and $\phi \to (1/y_0)\sqrt{6}$. Thus we have obtained a regular configuration, a wormhole, with a throat (minimum of r) at $y = \sqrt{y_0 y_1}$. This solution is *the only nonsingular solution among those with linear fields*. Wormholes will be discussed in more detail in Chapter 5.

Lastly, in the case $h = |C|$, $y_0 = 0$ (i.e., $\phi \to 0$ as $y \to \infty$) the sphere $y = 0$ turns out to be a horizon ($r \to r_0 = m = hy_1$, $e^\gamma \to 0$). In this case it is helpful to pass on to the curvature coordinates: $r = h(y + y_1)$, whence

$$ds^2 = (1 - m/r)^2 dt^2 - (1 - m/r)^{-2} dr^2 - r^2 d\Omega^2;$$

$$F^{10} = F_{01} = q/r^2; \qquad \phi = C/(r - m). \tag{4.34}$$

This solution is known as a BH with a scalar charge [30, 41, 50]. The metric has the same form as that of the Reissner–Nordström extremal BH ($m^2 = q^2$), to which this solution reduces in the case $C = 0$. It is noteworthy that, even though the ϕ field itself is singular at the horizon, its SET is finite there (it follows from regularity of the metric and the Einstein equations and can also be verified by a direct calculation).

A subsequent analysis has shown that the BHs (4.34) are unstable under small radial perturbations [73].

Solutions with nonconformal coupling

Vacuum and electrovacuum solutions have also been considered for a more general nonminimal coupling of a scalar field with gravity, corresponding to the Lagrangian (4.1) with

$$f(\phi) = 1 - \xi\phi^2, \quad \xi = \text{const} > 0;$$

$$h(\phi) \equiv 1, \quad U(\phi) \equiv 0. \tag{4.35}$$

At $\xi = 0$ and $\xi = 1/6$ this reduces to minimal and conformal couplings, respectively. For $\xi \neq 1/6$, it has been shown that the solution properties are to a large extent similar to those for $\xi = 1/6$: they are generically singular, but under some special relation between the integration constants, the singularity of the Einstein frame corresponds to a regular sphere in the Jordan frame [22, 69–71].

If $\xi > 1/6$, all solutions continued beyond this sphere have a second flat infinity, i.e., describe wormholes, but, as always in such cases, the region beyond the transition sphere is anti-gravitational in the sense that the effective gravitational constant is negative there. Similar wormhole solutions also exist for $\xi < 1/6$, but only under an additional inequality for integration constants, otherwise the conformally continued solution possesses a naked singularity [22, 71].

All these wormhole solutions with zero or small electric charges prove to be unstable under radial perturbations [69–71], and the characteristic time of perturbation growth is of the order of the time needed for a light signal to cover a distance of the order of the throat radius. Meanwhile, numerical calculations have shown [71] that the perturbation growth increment diminishes as the charge grows. It therefore seems possible that some values of the electric charge can lead to a stable wormhole.

BH solutions like (4.34) are absent at $\xi \neq 1/6$.

4.1.4 Anomalous (phantom) fields. The anti-Fisher solution

In the case of anomalous fields, both minimally and conformally coupled, there are more variants of the solution behavior. In particular, wormholes are obtained with both $\phi^{\rm min}$ and $\phi^{\rm conf}$ not in exceptional cases (as the metric (4.33)), but in the general case for $k<0$. From a formal viewpoint, this is connected with the fact that in the solutions (4.24)–(4.26) one can obtain either $k<0$ at $h \geq 0$ or $|k| > |h|$ if both $h<0$ and $k<0$. In these cases, at appropriate values of the constant u_1, the solution is defined on the interval $0<u<\pi/|k|$, and at both its ends there is a flat spatial infinity. From a physical viewpoint, the emergence of wormholes is connected with violation of the Null Energy Condition $T_{\mu\nu}\xi^\mu\xi^\nu \geq 0$, $\forall \xi^\mu : \xi^\nu\xi_\mu = 0$ — see Chapter 5.

As an example, let us consider the properties of the scalar-vacuum solution (4.30) for $\varepsilon = -1$, the phantom analogue of Fisher's solution, which, by analogy with "anti-de Sitter", we suggest to call "the anti-Fisher solution".

Branch A: $k > 0$. In this case it is helpful to pass over to the quasiglobal coordinate ρ by the transformation

$$\mathrm{e}^{-2ku} = 1 - 2k/\rho \equiv P(\rho), \tag{4.36}$$

and the solution takes the form

$$ds^2 = P^a dt^2 - P^{-a} d\rho^2 - P^{1-a} \rho^2 d\Omega^2,$$
$$\phi = -\frac{C}{2k} \ln P, \tag{4.37}$$

with the constants related by

$$a = h/k, \quad a^2 = 1 + C^2/(2k^2). \tag{4.38}$$

In the case $h < 0$, i.e., $m < 0$, we have $a < -1$, and, just as in the Fisher solution, a repulsive central singularity at $\rho = 2k$ corresponding to $u = \infty$.

The situation is, however, drastically different for $m > 0$, or $a > 1$. Indeed, the spherical radius r then has a finite minimum at $\rho = \rho_{\rm th} = (a+1)k$, corresponding to a throat of the size

$$r(\rho_{\rm th}) = r_{\rm th} = k(a+1)^{(a+1)/2}(a-1)^{(1-a)/2}, \tag{4.39}$$

and tends to infinity as $\rho \to 2k$. Moreover, for $a = 2, 3, \ldots$ the metric exhibits a horizon of order a at $\rho = 2k$ and admits a continuation to smaller ρ. A peculiar feature of such horizons is their infinite area. Such configurations have been named "cold black holes" (CBHs) [58, 59] since all of them have zero Hawking temperature.

Furthermore, all Kretschmann scalar constituents K_i [see (3.7)] behave as P^{a-2} as $\rho \to 2k$ and $P \to 0$. An exception is the value $a = 1$, in which case $C = 0$, $\phi = 0$, and the Schwarzschild solution is reproduced. Hence, at $\rho = 2k$ the metric has a curvature singularity if $a < 2$ (except for $a = 1$), a finite curvature if $a = 1$ or $a = 2$ and zero curvature if $a > 2$.

For noninteger $a > 2$, the qualitative behavior of the metric as $\rho \to 2k$ is the same as that near a horizon of infinite area, but a continuation beyond it is impossible due to nonanalyticity of the function $(1 - 2k/\rho)^a$ at $\rho = 2k$. Since geodesics terminate there at a finite value of the affine parameter, this is a space-time singularity (a singular horizon, as it has been called in Chapter 3) even though the curvature invariants tend there to zero.

Branch B: $k = 0$: the solution is defined in the range $u \in \mathbb{R}_+$ and is rewritten in terms of the quasiglobal coordinate $\rho = 1/u$ as follows:

$$ds^2 = \mathrm{e}^{-2h/\rho} dt^2 - \mathrm{e}^{2h/\rho}[d\rho^2 + \rho^2 d\Omega^2], \quad \phi = C/\rho. \tag{4.40}$$

As before, $\rho = \infty$ is a flat infinity, while at the other extreme $\rho \to 0$, the behavior is different for positive and negative mass $m = h/G$. Thus, for $m < 0$, $\rho = 0$ is a singular centre ($r = 0$ and all K_i are infinite). On the contrary, for $m > 0$, $r \to \infty$, and all $K_i \to 0$ as $\rho \to 0$. It is again a singular horizon: despite the vanishing curvature, the nonanalyticity of the metric in terms of ρ makes its continuation impossible.

Branch C: $k < 0$: the solution describes a wormhole with two flat asymptotes at $u = 0$ and $u = \pi/|k|$. The metric has the form

$$\begin{aligned} ds^2 &= \mathrm{e}^{-2hu} dt^2 - \frac{k^2\, \mathrm{e}^{2hu}}{\sin^2(ku)} \left[\frac{k^2\, du^2}{\sin^2(ku)} + d\Omega^2 \right] \\ &= \mathrm{e}^{-2hu} dt^2 - \mathrm{e}^{2hu} [d\rho^2 + (k^2 + \rho^2) d\Omega^2], \end{aligned} \tag{4.41}$$

where u is expressed in terms of the quasiglobal coordinate ρ, defined on the whole real axis $\mathbb{R}$, by $|k|u = \cot^{-1}(\rho/|k|)$. This metric is further discussed in Chapter 5.

4.1.5 Cold black holes in the anti-Fisher solution

Among different branches of the anti-Fisher solution, of greatest interest for us is the case of cold BHs. Let us briefly discuss their structure and properties.

For odd a, the principal geometric and causal properties, including the Carter-Penrose diagram, coincide with those of the Schwarzschild metric. Thus, at $\rho < 2k$, ρ is a temporal coordinate, t spatial, and the space-time is homogeneous and anisotropic, corresponding to the Kantowski-Sachs type of anisotropic cosmologies. The singularity at $\rho = 0$ $(r = 0)$ is spacelike (cosmological) and is reached by all timelike geodesics in a finite time interval after crossing the horizon.

For even a, the Penrose diagram is the same as that of the extreme Reissner-Nordström space-time; however, the physical meaning of the regions where $\rho < 2k$ is quite different. Since g_{22} and g_{33} change their signs at the horizon, the metric at $\rho < 2k$ has the signature $(-+++)$ instead of $(+---)$ at large ρ. The Lorentzian nature of space-time is still preserved, and one can verify that all geodesics are continued smoothly from one region to the other (the geodesic equations depend only on the Christoffel symbols and are invariant under the anti-isometry $g_{\mu\nu} \to -g_{\mu\nu}$). The time coordinate in that region is ρ since $g^{\rho\rho} < 0$ while the other diagonal components of $g_{\mu\nu}$ are positive. Thus, just as for odd a, we have a Kantowski-Sachs type cosmology with a spacelike singularity at $\rho = 0$ $(r = 0)$. The direction of the arrow of time can be arbitrary there since timelike geodesics that penetrate from the static region become spacelike (one cannot say for them where is the past and where is the future) there, and can even avoid the singularity.

The properties of the scalar field are no less exotic. According to (4.37), $\phi \to \infty$ as $\rho \to 2k$; this, however, does not contradict the regularity of the surface $\rho = 2k$ for $a \geq 2$ since the energy density

$$T_0^0 = -\frac{1}{2}A\phi'^2 = -\frac{C^2}{2}\frac{(\rho-2k)^{a-2}}{\rho^{a+2}}, \tag{4.42}$$

as well as the other components of T_μ^ν, are finite there (recall that for $a < 2$ the curvature invariants also diverge, together with T_μ^ν). Thus the infinite value of ϕ does not prevent the continuation of the space-time manifold to smaller ρ, where the solution is valid with $\phi = -(C \ln|P|)/(2k)$. On the other hand, the total scalar field energy, calculated as the conserved quantity corresponding to the timelike Killing vector, turns out to be infinite in the static region independently of a,

$$\mathcal{E} = \int T_0^0 \sqrt{g}\, d^3x = -2\pi C^2 \int \frac{d\rho}{\rho(\rho-2k)}, \tag{4.43}$$

and the integral logarithmically diverges at $\rho = 2k$. The divergence is related to the infinite spatial volume: the integral $\int \sqrt{{}^3g}\, d^3x$ diverges near $\rho = 2k$ even stronger than (4.43).

4.1.6 Vacuum and electrovacuum in Brans–Dicke theory

All static, spherically symmetric solutions of the Brans–Dicke (BD) theory (4.1), (4.7) with $U(\phi) \equiv 0$ are easily obtained from those described in Sec. 4.1.2 using the transformation (4.2). Namely, in the general theory (4.1) one can choose the Brans-Dicke scalar field parametrization $f(\phi) = \phi$, $h(\phi) = \omega(\phi)/\phi$, so that

$$L_{\rm STT} = \phi\tilde{R} + \frac{\omega(\phi)}{\phi}\tilde{g}^{\mu\nu}\phi_{,\mu}\phi_{,\nu} - 2U(\phi) + L_m, \tag{4.44}$$

and the transformation (4.2) reads

$$\tilde{g}_{\mu\nu} = \frac{1}{\phi}g_{\mu\nu}, \quad \frac{d\phi}{d\psi} = \phi|\omega(\phi) + 3/2|^{-1/2}. \tag{4.45}$$

where the metrics $\widetilde{g}_{\mu\nu}$ and $g_{\mu\nu}$ correspond to the Jordan and Einstein frames, respectively. In BD theory, $\omega = \text{const}$, so the fields ϕ and ψ are related by

$$\phi = \mathrm{e}^{\psi/\sqrt{|\omega+3/2|}}. \tag{4.46}$$

Recall that the BD theory is normal for $\omega > -3/2$, phantom for $\omega < -3/2$ and has no dynamics at $\omega = -3/2$; accordingly, Eq. (4.46) is meaningless at $\omega = -3/2$.

Substituting the metrics from Sec. 4.1.2 for $\widetilde{g}_{\mu\nu}$ and the corresponding expressions for ψ into (4.45) and (4.46), we obtain scalar-(electro)vacuum solutions of the BD theory. Their properties are analyzed in detail in quite a number of papers, see, e.g., [51, 58, 59, 85] and references therein. One can note that in the scalar-vacuum case ($q = 0$), the conformal mapping (4.45) with $\omega = \text{const}$ applied to Fisher's solution and the three branches of the anti-Fisher solution leads to Brans's four classes of vacuum BD solutions [48] (in other notations). The main conclusions are as follows.

1. In the case of a normal BD field, i.e., $\omega + 3/2 > 0$, all solutions possess naked singularities, with the only exception: a special wormhole solution existing at $\omega = 0$ [51] due to a conformal continuation, similar to that described in Sec. 4.1.3.
2. One of the generic types of BD solutions for all $\omega < -3/2$ is a wormhole with two asymptotically flat infinities.
3. Other generic types of BD solutions for all $\omega < -3/2$ contain singular horizons with an infinite area; their singular nature is connected with violation of analyticity rather than an infinite curvature. Some of these singularities are remote, i.e., correspond to infinite values of the canonical parameter along geodesics that reach them. In such cases the space-time is geodesically complete.
4. There is a discrete family of solutions with $\omega < -3/2$, parametrized by two integers, where the horizon of infinite area is regular and admits an extension beyond it (see diagrams 2b and 2d in Fig. 4.4). These configurations have been termed cold BHs since they inevitably have a zero Hawking temperature. Many of these configurations do not contain any singularities.
5. The electric charge adds some branches of solutions as compared with the vacuum case but does not drastically change the situation with singularities, BHs and wormholes [59].

4.1.7 Summary for massless scalar fields

From our consideration of exact static solutions with massless scalar fields, one can draw the following basic conclusions concerning BHs.

1. In GR with a normal scalar field $\phi^{\rm min}$ ($\varepsilon = +1$), there are no solutions with horizons.

2. In GR with a phantom scalar field $\phi^{\rm min}$ ($\varepsilon = -1$), there are "cold BH" solutions containing horizons of infinite area under special discrete values of the parameters (integration constants). There are no horizons of finite area.

3. In GR with a conformal scalar field $\phi^{\rm conf}$, with the normal sign of kinetic energy ($\varepsilon = +1$), there is an unstable special solution in the form of an extremal BH (with a second-order horizon).

4. Other kinds of nonminimally coupled scalar fields, scalar-tensor theories: at $\varepsilon = +1$ cases like item three are possible in principle but are yet unknown; at $\varepsilon = -1$ the situation is similar to item two. For instance, such discrete families of electrically neutral and charged cold BHs have been found in the Brans-Dicke theory. Some of them are globally regular.

4.2 Scalar fields with arbitrary potentials. No-go theorems

4.2.1 What is the use of no-go theorems?

In most of the applications, it is necessary to consider more general scalar fields than those discussed in the previous section, namely, fields with nonzero potentials as well as various interactions, generalized or modified scalar field theories. Meanwhile the equations for such self-gravitating scalar fields can be explicitly integrated only in very rare special cases even for fields with high symmetry, such as isotropic models in cosmology and static, spherically symmetric systems. Therefore, of great value are general statements or theorems about the properties of such systems which can be obtained without completely solving the field equations.

For static, spherically symmetric scalar-vacuum configurations in GR, described by the action

$$S = \int d^4x \sqrt{|g|}\,\bigl[R/\varkappa^2 + L_{\rm sc}\bigr],$$
$$L_{\rm sc} = g^{\mu\nu}\phi_{,\mu}\phi_{,\nu} - 2V(\phi), \qquad (4.47)$$

where R is the scalar curvature and $\varkappa^2$ the gravitational constant, the following theorems are known:

A. "No-hair" theorems ([5, 31]), claiming that asymptotically flat black holes with a nontrivial scalar field cannot exist under certain conditions imposed on the potential $V(\phi)$.

B. The generalized Rosen theorem, claiming that a particle-like solution (PLS) (that is, an asymptotically flat solution with a regular centre) cannot exist if $V \geq 0$ [84].

C. The nonexistence theorem for regular solutions without a centre (e.g., wormhole solutions) [53].

D. The global structure theorem [53], claiming that the list of possible global causal structures (and the corresponding Carter-Penrose diagrams) for configurations with arbitrary potentials $V(\phi)$ and any asymptotic behavior is the same as the analogous list in the case $\phi = \text{const}$, namely it includes the Minkowski (or AdS), Schwarzschild, de Sitter and Schwarzschild-de Sitter structures.

A number of known exact solutions represent examples of configuration admitted by these theorems: black hole solutions with scalar fields where $V(\phi) \geq 0$ but without asymptotic flatness; asymptotically flat PLS and black holes with at least partly negative potentials $V(\phi)$, etc. All this, taken together, leads to a clear enough picture of what can and what cannot be expected from minimally coupled scalar fields in GR. A detailed description of all these results is beyond the scope of our presentation here, see, e.g., [53, 84] and references therein.

There are quite a number of generalizations of the action (4.47), for which it is of interest to know whether or not the above Statements A–D hold, and if yes, under which additional conditions. In particular:

1. Multidimensional spherically symmetric configurations in GR, with the scalar field Lagrangian given in (4.47) and the metric

$$ds^2 = \mathrm{e}^{2\gamma} dt^2 - \mathrm{e}^{2\alpha} du^2 - \mathrm{e}^{2\beta} d\Omega_{d_0}^2, \tag{4.48}$$

where α, β, γ are functions of the radial coordinate u and $d\Omega_{d_0}^2$ is a linear element on a d_0-dimensional sphere of unit radius.

2. More general scalar field Lagrangians in GR, such as $L_{\rm sc} = F(\phi,\ I)$, where $I = g^{\mu\nu}\phi_{,\mu}\phi_{,\nu}$ and F is an arbitrary function of two variables (the so-called generalized k-essence).

3. Sets of scalar fields $\vec{\phi} = \{\phi^K\}$, $K = \overline{1, N}$ of σ model type, considered as coordinates in the N-dimensional target space $\mathbb{T}_\phi$, so that

$$L_{\rm sc} = H_{KL}(\vec{\phi})(\partial\phi^K,\ \partial\phi^L) - 2V(\vec{\phi}), \tag{4.49}$$

where the target-space metric H_{KL} (usually supposed to be positive-definite) and the potential V are functions of the N variables ϕ^K, and we use the notation

$$(\partial y,\ \partial z) = g^{\mu\nu}\partial_\mu y \partial_\nu z. \tag{4.50}$$

4. Scalar-tensor theories (STT) of gravity, with the D-dimensional action

$$S_{\rm CTT} = \int d^D x \sqrt{g}[f(\phi)\mathcal{R} + h(\phi)(\partial\phi)^2 - 2U(\phi)], \qquad (4.51)$$

where $(d\phi)^2 = (\partial\phi, \partial\phi)$ and f, h, U are arbitrary functions of the scalar field.

5. Theories of gravity with higher derivatives (curvature-nonlinear theories) where the scalar curvature R in (4.47) is replaced by a function $f(\mathcal{R})$, or, more generally, including, for example, the invariant $R_{\mu\nu}R^{\mu\nu}$.

6. More general multidimensional configurations such as products like

$$\mathbb{M}^D = \mathbb{R}_u \times \mathbb{M}_0 \times \mathbb{M}_1 \times \mathbb{M}_2 \times \cdots \times \mathbb{M}_n, \qquad (4.52)$$

where $\mathbb{M}_{\rm ext} = \mathbb{R}_u \times \mathbb{M}_0 \times \mathbb{M}_1$ is the "external" manifold with the metric (4.48), $\mathbb{R}_u \subseteq \mathbb{R}$ is the range of the radial coordinate u, $\mathbb{M}_1$ is the time axis, and $\mathbb{M}_0 = \mathbb{S}^{d_0}$. Furthermore, $\mathbb{M}_2, \ldots, \mathbb{M}_n$ are the "internal" factor spaces of arbitrary dimensions d_i, $i = 2, \ldots, n$; by this notation, $\dim \mathbb{M}_0 = d_0$ and $\dim \mathbb{M}_1 = d_1 = 1$.

This list can be continued and/or its items can be combined to obtain new generalizations.

We will consider here static, spherically symmetric systems in GR, in space-times of arbitrary dimension D with a sufficiently general σ-model source. The proofs turn out to be almost the same as for similar systems in four dimensions (4.47), but the present formulation is well suited for (multi-)scalar-tensors and multidimensional generalizations.

4.2.2 Basic equations

Consider D-dimensional GR with the set of scalar fields (4.49). As in four dimensions, the Einstein equations can be written in two equivalent forms:

$$G_\mu^\nu \stackrel{\rm def}{=} \mathcal{R}_\mu^\nu - \frac{1}{2}\delta_\mu^\nu \mathcal{R} = -\varkappa^2 T_\mu^\nu, \qquad (4.53)$$

or

$$\mathcal{R}_\mu^\nu = -\varkappa^2 \widetilde{T}_\mu^\nu \stackrel{\rm def}{=} -\varkappa^2 \left(T_\mu^\nu - \frac{\delta_\mu^\nu}{D-2} T_\alpha^\alpha \right), \qquad (4.54)$$

where T_μ^ν is the stress-energy tensor (SET). For the fields (4.49) it is given by

$$T_\mu^\nu = \partial^\nu \vec{\phi} \partial_\mu \vec{\phi} - \frac{1}{2}\delta_\mu^\nu L_{\rm sc}, \qquad (4.55)$$

or in the "tilded" version,

$$\widetilde{T}_\mu^\nu = \partial^\nu \vec{\phi} \partial_\mu \vec{\phi} - \delta_\mu^\nu V(\vec{\phi}), \tag{4.56}$$

where two arrows denote a scalar product in the target space $\mathbb{T}_\varphi$: $\vec{a}\vec{b} = H_{KL} a^K b^L$.

The static, spherically symmetric metric (4.48) is written with an arbitrary radial coordinate u. Now, it is convenient for our purposes to use the quasiglobal coordinate $u = \rho$ corresponding to the gauge condition $\alpha + \gamma = 0$, so that the metric takes the form

$$ds^2 = A(\rho)dt^2 - \frac{du^2}{A(\rho)} - r^2(\rho) d\Omega_{d_0}^2, \tag{4.57}$$

where we have denoted $r(\rho) = \mathrm{e}^\beta$ and $A(\rho) = \mathrm{e}^{2\gamma} \equiv \mathrm{e}^{-2\alpha}$. This choice is preferable for considering Killing horizons, described as zeros of the function $A(\rho)$. The reason is that in a close neighborhood of a horizon the coordinate ρ defined in this way varies (up to a positive constant factor) like manifestly well-behaved Kruskal-like coordinates used for an analytic continuation of the metric. This fact refers to the geometry of the (t, ρ) subspace and does not depend on the geometry of other dimensions and even on their number.

With this choice of the coordinate gauge, the nonzero components of the Ricci tensor for the metric (4.57) read

$$R_t^t = -\frac{1}{2r^{d_0}}(A' r^{d_0})', \tag{4.58}$$

$$R_\rho^\rho = -\frac{1}{2r^{d_0}}(A' r^{d_0})' - d_0 A \frac{r''}{r}, \tag{4.59}$$

$$R_\theta^\theta = \frac{d_0 - 1}{r^2} - A\left[\frac{r''}{r} + (d_0 - 1)\frac{r'^2}{r^2} + \frac{A'r'}{Ar}\right], \tag{4.60}$$

where θ is any of the spherical angles in (4.57). Accordingly, the scalar field equations and four different combinations of components of Eq. (4.54) can be written as follows:

$$[A r^{d_0} H_{KL} (\varphi^L)']' = r^{d_0} \partial V / \partial \varphi^K; \tag{4.61}$$

$$(A' r^{d_0})' = -(4/d_0) r^{d_0} \varkappa^2 V; \tag{4.62}$$

$$d_0 r''/r = -\varkappa^2 (\vec{\phi}')^2; \tag{4.63}$$

$$A(r^2)'' - r^2 A'' + (d_0 - 2) r' (2Ar' - A'r) = 2(d_0 - 1); \tag{4.64}$$

$$d_0(d_0 - 1)(1 - A{r'}^2) - d_0 A' r r' = -A r^2 (\vec{\phi}')^2 + 2r^2 V. \tag{4.65}$$

Eqs. (4.62), (4.63) and (4.64) are the components $\binom{t}{t}$, $\binom{t}{t}-\binom{\rho}{\rho}$ and $\binom{t}{t}-\binom{\theta}{\theta}$ of (4.54) respectively, and (4.65) is the $\binom{\rho}{\rho}$ component of (4.53). We have written $(N+4)$ equations for $(N+2)$ unknowns φ^K, A and r, but there are only two independent equations among (4.62)–(4.65); in particular, (4.65) is a first integral of the other equations. So this set of equations is determined.

4.2.3 Global structure theorems

The first theorem concerns the nonexistence of wormholes, horns and flux tubes. A ***wormhole*** is, by definition, a configuration with two asymptotic regions that have large $r(\rho)$ (see Chapter 5), hence $r(\rho)$ must have at least one regular minimum. A ***flux tube*** is a configuration with $r = \text{const} > 0$, a static (d_0+1)-dimensional cylinder. A ***horn*** is a geometry that tends to a flux tube at one of its asymptotics, which happens if $r(\rho) \to \text{const} > 0$ at one of the ends of the range of ρ. Such "horned particles" with a flat asymptotic region have been discussed as possible remnants of black hole evaporation [19].

Theorem 4.1 (on configurations without a centre). *Eqs.* (4.61)–(4.65) *for* $D \geq 4$ *and positive-definite* H_{KL} *do not admit* (i) *solutions where the function* $r(\rho)$ *has a regular minimum*, (ii) *solutions describing a horn*, *and* (iii) *flux-tube solutions with* $\varphi^K \neq \text{const}$.

A proof rests on Eq. (4.63), implying $r'' \leq 0$, which actually expresses the null energy condition valid for the SET T_μ^ν as long as the matrix H_{KL} is positive-definite. As a result, not only are wormholes as global entities impossible but even wormhole throats.

Another theorem concerns the possible number and order of Killing horizons, coinciding with the number and order of zeros of $A(\rho)$. A simple (first-order) or any odd-order horizon separates a static region, $A > 0$ (R-region), from a nonstatic region, $A < 0$ where (4.48) is a homogeneous cosmological Kantowski–Sachs-type metric (a T-region). A horizon of even order separates regions with the same sign of $A(\rho)$.

The disposition of horizons unambiguously determines the global causal structure of space-time (up to identification of isometric surfaces, if any). The following theorem severely restricts such possible dispositions.

Theorem 4.2 (on the global structure). *Consider solutions to Eqs.* (4.61)–(4.64) *for* $D \geq 4$. *Let there be a static region* $a < \rho < b \leq \infty$.

Then:

(i) *all horizons are simple;*
(ii) *no horizons exist at* $\rho < a$ *and at* $\rho > b$.

A proof of this theorem [53, 84] employs the properties of Eq. (4.64), which can be rewritten in the form

$$r^4 B'' + (d_0 + 2) r^3 r' B' = -2(d_0 - 1), \tag{4.66}$$

where $B(\rho) = A/r^2$. At points where $B' = 0$, we have $B'' < 0$. Therefore $B(\rho)$ cannot have a regular minimum. Having once become negative while moving to the left or to the right along the ρ axis, $B(\rho)$ (and hence $A(\rho)$) cannot return to zero or positive values.

By Theorem 4.2, there can be at most two simple horizons around a static region. A second-order horizon separating two nonstatic regions can appear, but this horizon is then unique, and the model has no static region.

The possible dispositions of zeros of the function $A(\rho)$, and hence the list of possible global causal structures, are thus the same as in the case of vacuum with a cosmological constant. The latter is a solution to Eqs. (4.61)–(4.64) with $\varphi^K = \text{const}$, $\varkappa^2 V = \Lambda = \text{const}$, and the metric

$$ds^2 = A(r)dt^2 - \frac{dr^2}{A(r)} - r^2\, d\Omega^2_{d_0}, \tag{4.67}$$

$$A(r) = 1 - \frac{2m}{(d_0 - 1) r^{d_0 - 1}} - \frac{2\Lambda r^2}{d_0(d_0 + 1)}. \tag{4.68}$$

This is the multidimensional Schwarzschild-de Sitter (or Tangherlini-de Sitter) solution. Its special cases correspond to the Schwarzschild ($d_0 = 2$) and Tangherlini [395] ($d_0 \geq 2$) solutions[15] when $\Lambda = 0$ and the de Sitter solution in arbitrary dimensions when $m = 0$, called anti-de Sitter (AdS) in case $\Lambda < 0$. For $\Lambda > 0$, if m is positive but smaller than the critical value

$$m_{\rm cr} = \frac{d_0 - 1}{d_0 + 1} \left[\frac{d_0(d_0 - 1)}{2\Lambda}\right]^{(d_0 - 1)/2}, \tag{4.69}$$

there are two horizons, the one at smaller r being interpreted as a black hole horizon and the other as a cosmological horizon. At $m = m_{\rm cr}$, the two

[15]The mass M in conventional units, say, grams, is obtained by writing $m = GM$ where G is a $(d_0 + 2)$-dimensional analogue of Newton's constant. The coefficient of m is chosen in (4.70) and accordingly in (4.68) in such a way that at large r in the case $\Lambda = 0$, when the space-time is asymptotically flat, a test particle at rest experiences a Newtonian acceleration equal to $-GM/r^{d_0}$.

horizons merge, and there are two homogeneous nonstatic regions separated by a double horizon. The solution with $m > m_{\rm cr}$ is purely cosmological and has no Killing horizon. In the cases $m < 0$ and/or $\Lambda < 0$ there is at most one simple horizon.

Possible behaviors of $A(r)$ and the corresponding Carter–Penrose diagrams are presented in Figs. 4.1 and 4.2.

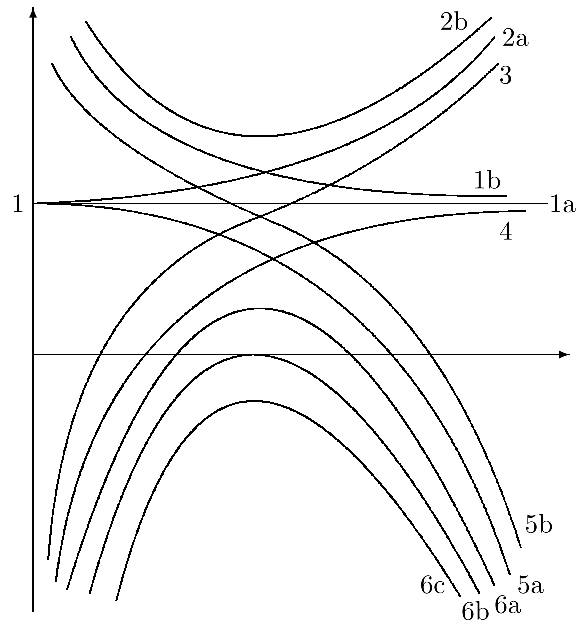


Figure 4.1 The behavior of $A(r)$, Eq. (4.68), for different values of m and Λ.

In (2+1)-dimensional gravity ($d_0 = 1$) we have a still shorter list of global structures: at most one simple horizon is possible.

Theorems 1 and 2 are independent of the form of the potential and of any assumptions about spatial asymptotics.

4.2.4 No-hair theorem

Let us now consider asymptotically flat space-times, which means, in terms of the metric (4.57), that without loss of generality, $r \approx \rho$ and the function

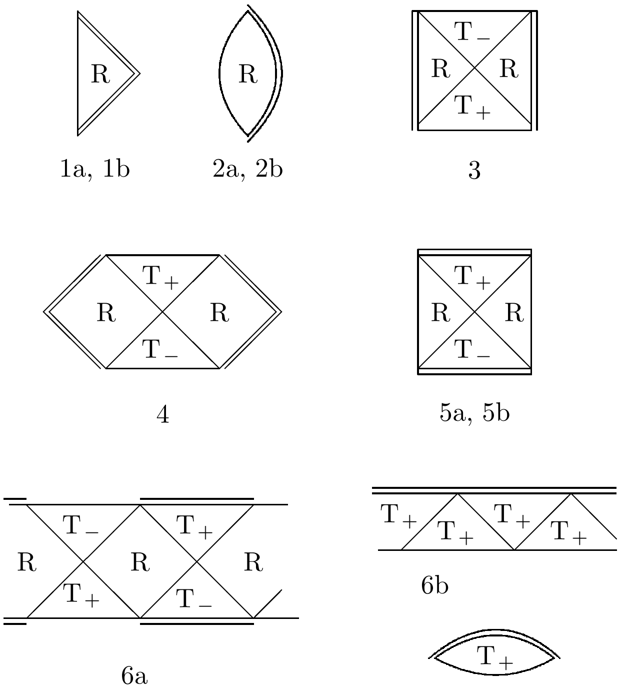


Figure 4.2 Carter–Penrose diagrams for different cases of the metric (4.67) and (4.68), labelled according to Fig. 4.2.3. The R and T letters correspond to R and T space-time regions while T_+ and T_- denote expanding and contracting T region (i.e., with r increasing and decreasing with time, respectively). Single lines on the border of the diagrams denote $r = 0$, double lines $r = \infty$. Diagrams 6b and 6c are drawn for the case of expanding Kantowski-Sachs cosmologies; to obtain diagrams for contracting models, one should merely interchange $r = 0$ and $r = \infty$ and replace T_+ with T_-. Diagrams 6a and 6b admit identification of isometric timelike sections.

$A(\rho) \approx A(r)$ has the Tangherlini form, (i.e., 4.68) with $\Lambda = 0$:

$$A(r) = 1 - \frac{2m}{(d_0 - 1) r^{d_0 - 1}}, \tag{4.70}$$

as $\rho \to \infty$. It then follows from the field equations that the SET components, and hence the quantities V and $(\vec{\phi})^2$, decay at large $\rho \approx r$ quicker than $r^{-(d_0+1)}$.

Let us now prove the following no-hair theorem [68], extending to our system the theorems known in four dimensions [5, 31]:

Theorem 4.3 (on external scalar fields of black holes). *Given Eqs.* (4.61)–(4.65) *for* $D \geq 4$, *with a positive-definite matrix* $H_{KL}(\vec{\phi})$ *and* $V(\vec{\phi}) \geq 0$, *the only asymptotically flat black hole solution is characterized by* $V \equiv 0$, $\vec{\phi} = \text{const}$ *and the Tangherlini metric* (4.67), (4.70) *in the whole range* $h < \rho < \infty$ (*the domain of outer communication of the black hole*) *where* $\rho = h$ *is the event horizon.*

At the event horizon $\rho = h$ we have by definition $A = A(h) = 0$, and $A > 0$ at $\rho > h$. By Theorem 4.2, the horizon should be simple, so that $A \sim \rho - h$ as $\rho \to h$. Consider the function

$$\mathcal{F}_1(\rho) = \frac{r^{d_0}}{r'}[2V - A\vec{\phi}'^2]. \tag{4.71}$$

One can verify that

$$\mathcal{F}'(\rho) = \mathcal{F}_2(\rho) \stackrel{\text{def}}{=} r^{d_0 - 1}\left[2d_0 V + (d_0 - 1)\frac{\vec{\phi}'^2}{r'^2} + A\vec{\phi}'^2\right]. \tag{4.72}$$

To do so, when calculating $\mathcal{F}_1'$, one should substitute $\vec{\phi}''$ from (4.61), r'' from (4.63) and A' from (4.65). Let us integrate (4.72) from h to infinity:

$$\mathcal{F}_1(\infty) - \mathcal{F}_1(h) = \int_h^\infty \mathcal{F}_2(\rho)\, d\rho. \tag{4.73}$$

Since $r'(\infty) = 1$ and $r'' \leq 0$, we have $r' > 1$ in the whole range of ρ, but $r'(h) < \infty$. Indeed, regularity of the horizon implies a finite value of the Kretschmann scalar given by (3.7), hence finite values of all its constituents. Of interest for us is the condition $|A'r'| < \infty$. Since $A'(h) > 0$ (the horizon is simple due to the Global structure theorem), its finiteness means $r'(h) < \infty$.[16]

The quantity $\mathcal{F}_1(h)$ should be finite, since otherwise we would have either V or $A\vec{\phi}'^2$ infinite, leading to infinite SET components (see (4.55)) and, via the Einstein equations, to a curvature singularity.

If, however, we admit a nonzero value of $A\vec{\phi}'^2$ at $\rho = h$, the integral in (4.73) will diverge at $\rho = h$ due to the second term in brackets in (4.72),

[16]The expressions (3.7) are written for the case $D = 4$, but for the metric (4.57) they have the same form up to numerical factors.

and this in turn leads to an infinite value of $\mathcal{F}_1(h)$. Therefore $A\vec{\phi}'^2 \to 0$ as $\rho \to h$, and we conclude that

$$\mathcal{F}_1(h) = \frac{2r^{d_0}(h)}{r'(h)} V(h) \geq 0.$$

On the other hand, $\mathcal{F}_1(\infty) = 0$ due to the asymptotic flatness conditions. Thus, in Eq. (4.73) there is a nonpositive quantity in the left-hand side and a nonnegative quantity on the right. The only way to satisfy (4.73) is to put $V \equiv 0$ and $\vec{\phi}' \equiv 0$ in the whole range $\rho > h$, and the only solution for the metric then has the Tangherlini form. □

As follows from the scalar field equations (4.61), the equality $V = 0$ should take place where $\partial V/\partial\varphi^K = 0$, i.e., at an extremum or saddle point of the potential, and it should obviously be a minimum for stable equilibrium.

It is of interest that one of the key points of the above proof, that $A\vec{\phi}'^2 = 0$ at $\rho = h$, might be obtained from smoothness considerations. Indeed, since $A \sim \rho - h$ near $\rho = h$, a nonzero value of $A\vec{\phi}^2$ means that some of $(\phi^K)'$ behave as $(\rho - h)^{-1/2}$, violating the C^1 requirement for the scalar fields. Our proof is "more economical" since it only uses the requirement of space-time regularity at the horizon.

One can also note that our no-hair theorem is in a complementarity relation with a recent black hole uniqueness theorem [348] (see [349] for recent generalizations and [201, 290] for earlier reviews). In D-dimensional general relativity coupled to the σ-model (4.49) with $V \equiv 0$, it has been proved without assuming spherical symmetry at the outset that "the only black hole solution with a regular, nonrotating event horizon in an asymptotically flat, strictly stationary domain of outer communication is the Schwarzschild-Tangherlini solution with a constant mapping ϕ" [348]. In contrast to that, our Theorem 4.3 applies to σ-models with arbitrary $V(\vec{\phi}) \geq 0$ but selects the Tangherlini solution among spherically symmetric configurations.

There are a number of other theorems of similar nature whose common meaning is that any stationary (rather than static or spherically symmetric only) black hole is completely and uniquely characterized by solutions to the Einstein equations determined by the parameters M (mass), Q (electric [and magnetic] charge) and J (angular moment). It is this statement that is meant in Wheeler's claim [424] that "Black holes have no hair".

There exist many generalizations and analogues of such theorems in extensions of GR in diverse dimensions as well as for some non-asymptotically flat configurations (see, e.g., [31, 161, 290] for reviews).

4.2.5 Two expressions for the mass and the properties of particle-like solutions

In this subsection we will discuss particle-like solutions, i.e., solutions with a flat asymptotic and a regular centre. We begin by deriving two general expressions for the active gravitational (Tangherlini) mass m of a D-dimensional configuration with the metric (4.48) and an arbitrary SET compatible with the regular centre and asymptotic flatness conditions.

One expression is easily obtained from the $\binom{t}{t}$ component of Eq. (4.53) which may be written in the curvature coordinates ($u = r$) in the following way:

$$\frac{d_0}{(d_0-1)r^{d_0}}\frac{dm}{dr} = \varkappa^2 T_t^t, \tag{4.74}$$

where $m(r)$ is the mass function,

$$m(r) \stackrel{\text{def}}{=} \frac{d_0-1}{2} r^{d_0-1}(1-\mathrm{e}^{-2\alpha}), \tag{4.75}$$

generalizing the well-known 4-dimensional mass function $m(r) = \frac{1}{2}r(1-\mathrm{e}^{-2\alpha})$. For a system with a regular centre ($r=0$), the function $m(r)$, expressed from (4.74) as

$$m(r) = \frac{d_0-1}{d_0}\varkappa^2 \int_0^r T_t^t r^{d_0} dr, \tag{4.76}$$

can be interpreted as the mass inside a sphere of radius r. If, in addition, the space-time is asymptotically flat, this integral converges at large r and if taken from zero to infinity, gives the full Tangherlini mass $m = m(\infty)$. The constant $\varkappa^2$ is expressed in terms of d_0 and the multidimensional Newtonian constant G (such that $m = GM$, see footnote 15) if we require the validity of the usual expression for mass in terms of density, $M = \int T_t^t\, dv$ (dv being the element of volume) in the flat space limit. One thus obtains

$$\varkappa^2 = \frac{d_0}{d_0-1}s(d_0)G,$$

$$s(d_0) = 2\pi^{(d_0+1)/2}/\Gamma\left(\frac{d_0+1}{2}\right), \tag{4.77}$$

where $s(d_0)$ is the area of a d_0-dimensional sphere of unit radius and Γ is Euler's gamma function. In case $D=4$ we have, as usual, $\varkappa^2 = 8\pi G$.

Eq. (4.76) for the Tangherlini mass is easily rewritten in terms of any radial coordinate, e.g., the quasiglobal ρ coordinate used in Eqs. (4.61)–(4.65):

$$m = \frac{d_0 - 1}{d_0} \varkappa^2 \int_{\rho_c}^{\infty} T_t^t(\rho) r^{d_0} r' d\rho, \tag{4.78}$$

where ρ_c is the value of ρ at the centre.

On the other hand, one can integrate the $\binom{t}{t}$ component of Eq. (4.54), which, in terms of the same ρ coordinate for R_t^t, assumes the form

$$\frac{1}{2r^{d_0}} (A' r^{d_0})' = \varkappa^2 \widetilde{T}_t^t. \tag{4.79}$$

For an asymptotically flat metric (4.57) with a regular centre, integration of (4.79) over the whole range of ρ gives

$$m = \frac{\varkappa^2}{d_0} \int_{\rho_c}^{\infty} [(d_0 - 1) T_t^t - T_i^i] r^{d_0} d\rho, \tag{4.80}$$

where the index i enumerates spatial coordinates. It is a multidimensional analogue of Tolman's well-known formula [398] for the mass of a regular matter distribution in general relativity. Comparing the expressions (4.78) and (4.80), we obtain the following *universal identity valid for any particle-like static, spherically symmetric configuration in D-dimensional GR*:

$$\int_{\rho_c}^{\infty} [(r' - 1)(d_0 - 1) T_t^t + T_i^i] r^{d_0} d\rho = 0. \tag{4.81}$$

For the σ-model (4.49), Eq. (4.80) takes the form

$$m = -\frac{2\varkappa^2}{d_0} \int_{\rho_c}^{\infty} V(\vec{\phi}) r^{d_0} d\rho, \tag{4.82}$$

leading to a multidimensional version of what has been previously called the generalized Rosen theorem [84]: *a static, spherically symmetric particle-like solution with positive mass cannot be obtained with scalar fields having a nonnegative potential V.*

An even stronger no-go theorem follows from the universal identity (4.81) [68]:

Theorem 4.4 (on particle-like solutions). *Eqs.* (4.61)–(4.65) *with $D \geq 4$ for the σ model* (4.49) *do not admit any particle-like solution if the matrix H_{KL} is positive-definite and $V \geq 0$.*

In other words, even negative-mass particle-like solutions can only be obtained with (at least partly) negative potentials.

To prove the theorem, it is sufficient to show that the expression in brackets in (4.81) is positive for any nontrivial solution under the conditions of the theorem — it will then follow that the integral is positive, hence it is not a particle-like configuration. This expression is

$$\frac{1}{2}(d_0 - 1)A(\vec{\phi}')^2 + V[2 + (d_0 - 1)r'].$$

Its positivity is evident since, as already mentioned, $r' = 1$ at the flat asymptotic and, due to $r'' \leq 0$, we have $r' \geq 1$ in the whole range of ρ. □

4.3 Rotating black holes

Until now we have been discussing only spherically symmetric BHs, whose angular momentum is zero. Now we will very briefly describe the basic properties of rotating BHs. We will assume them to be electrically neutral both for simplicity and in connection with the small role of possible BH charges in astrophysical conditions.

Rotating BHs form as a result of the gravitational collapse of rotating bodies (practically all celestial bodies are rotating) and at BH mergers. Therefore rotating BHs are much more realistic than spherically symmetric ones. Their structure and properties are, however, much more complicated.

Stationary rotating BHs are described by the Kerr metric [229]

$$\begin{aligned} ds^2 = &\left(1 - \frac{2mr}{\rho^2}\right)dt^2 + \frac{4m\,r\,a^2\sin^2\theta}{\rho^2}d\phi\,dt - \frac{\rho^2}{\Delta}dr^2 - \rho^2 d\theta^2 \\ &-\left(r^2 + a^2 + \frac{2m\,r\,a^2}{\rho^2}\right)\sin^2\theta\,d\phi^2, \end{aligned} \tag{4.83}$$

where

$$a = J/m, \qquad \rho^2 = r^2 + a^2\cos^2\theta, \qquad \Delta = r^2 - 2m\,r + a^2. \tag{4.84}$$

Here, $(t,\ r,\ \theta,\ \phi)$ are the so-called Boyer–Lindquist coordinates [45] (initially, Kerr obtained his metric in less convenient notations). The metric (4.83) represents an asymptotically flat (as $r \to \infty$) stationary, axially symmetric solution to the Einstein equations in vacuum ($T^\nu_\mu \equiv 0$). The stationarity and axial symmetry are determined by the existence of the Killing vectors $\xi^\mu = (1, 0, 0, 0)$ and $\eta^\mu = (0, 0, 0, 1)$, respectively (with the usual order of enumerating the coordinates: $t,\ r,\ \theta,\ \phi$). The stationarity, instead of staticity, of the metric manifests itself in its noninvariance under the substitution $t \mapsto -t$.

The solution is parametrized by two constants, the mass m and the angular momentum J; at $J = 0$ it turns into the Schwarzschild solution.

A true curvature singularity occurs at $\rho = 0$, $r = 0$, $\theta = \pi/2$, and an inspection shows that this singularity has the shape of a ring (this fact is not evident in the Boyer-Lindquist coordinates and is established using another parametrization of the metric, see, e.g., [168, 296]). The surfaces where $\Delta = 0$, i.e.,

$$r_\pm = m \pm \sqrt{m^2 - a^2}, \tag{4.85}$$

are horizons. Evidently, the horizons exist if $m^2 \geq a^2$, and in the case of inequality there are two simple horizons while at equality there is only one extremal (double) horizon.

Consider the metric (4.83) with horizons. In this case, the norm of the Killing vector ξ^μ, equal to

$$\xi_\mu \xi^\mu = g_{tt} = \frac{\Delta - a^2 \sin^2\theta}{\rho^2}, \tag{4.86}$$

vanishes on the surface where

$$\Delta = a^2 \sin^2\theta, \tag{4.87}$$

i.e., outside the horizon (with the surface $\Delta = 0$, i.e., the horizon $r = r_+$, it has only two common points, namely, the poles $\theta = 0$ and $\theta = \pi$, see Fig. 4.3). The surface (4.87) is called the stationarity limit because

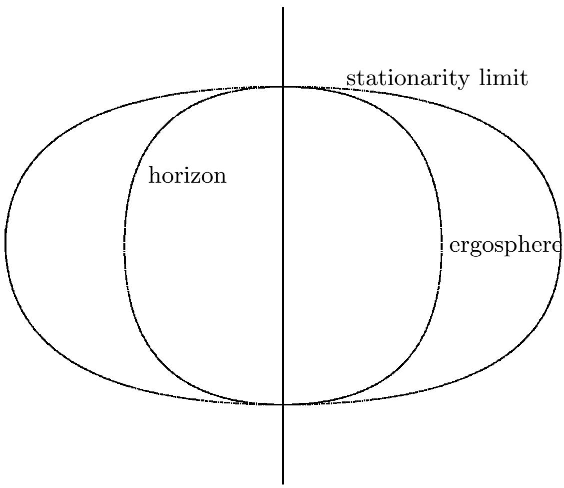


Figure 4.3 The horizon and the ergosphere of a Kerr BH, a meridional section.

stationary observers cannot exist on and inside this surface. Besides, the surface (4.87) is a surface of infinite redshift for photons emitted from it to spatial infinity.

The region of space between the horizon $r = r_+$ and the surface (4.87) has received the name of the ergosphere. This name is connected with the opportunity to extract the rotational energy of the Kerr BH with the aid of the Penrose process [325]. The point is that the ergosphere contains orbits of massive particles with a negative total energy (this means that the absolute value of the binding energy exceeds the particle mass). Then one can launch, say, a spacecraft into the ergosphere, where it will emit particles with negative energy and return to the usual spatial region with a larger energy than it initially had. As to particles with a negative energy, their motion is directed against the BH rotation and, being absorbed by the BH, they diminish its rotation rate and total rotational energy.

It has been shown [18, 183, 184, 434] that particle collisions close to horizons of rotating black holes can lead to extremely high energies of the order of the Grand Unification scale in the centre-of-mass reference frame of colliding particles. According to [18], close to the horizon of an extremal Kerr black hole, there is a resonance for the centre of mass (CM) energy of two scattering particles: a pole for some special value of the angular momentum of the scattered particle was found, showing that the centre of mass energy can be arbitrarily large. The papers [183, 184] show that the same effect can occur in nonextremal black holes if one takes into account the possibility of multiple scattering of a particle: at the first scattering close to the horizon the particle gets an angular momentum close to its critical value, and at the second scattering, due to the resonance effect, it can acquire an energy at the level of high energy physics, Grand Unification or even Planck values.

Assuming that a rotating black hole is located at the centre of an active galactic nucleus, the time needed for a particle to come from the accretion disk to the first scattering point was estimated [184] and turned out to be of the order of a week. Then, the time needed for a high-energy particle to leave the ergosphere to be observed on Earth as a particle of ultra-high energy cosmic rays (UHECR) was evaluated. This time was also found to be quite reasonable: of the order of days, plus the time needed to travel to the observer.

Electrically charged stationary rotating BHs are described by the Kerr-Newman solution to the Einstein–Maxwell equations, characterized by the metric in the Boyer–Lindquist coordinates similar to those in (4.83)

$$ds^2 = \left(1 - \frac{2mr - q^2}{\rho^2}\right) dt^2 + \frac{(2m\,r - q^2)2a^2 \sin^2\theta}{\rho^2} d\phi\, dt - \frac{\rho^2}{\Delta} dr^2 - \rho^2 d\theta^2$$
$$- \left(r^2 + a^2 + \frac{(2m\,r - q^2)a^2}{\rho^2}\right) \sin^2\theta\, d\phi^2, \qquad (4.88)$$

with

$$\rho^2 = r^2 + a^2\cos^2\theta, \qquad \Delta = r^2 - 2m\,r + a^2 + q^2, \qquad (4.89)$$

where q is the electric charge. m and a have the same meaning as in Kerr's solution. Evidently, ρ^2 is the same as in (4.84) while Δ has the new contribution q^2. The electromagnetic field of the BH is determined by the vector potential A_μ given by

$$A_\mu dx^\mu = -\frac{qr}{\rho^2}(dt - a\sin^2\theta\, d\phi). \qquad (4.90)$$

This solution reduces to the Reissner–Nordström solution if $a = 0$, and to the Kerr solution if $q = 0$. A detailed analysis of the properties of Kerr–Newman BHs can be found in the book [168].

At large charges and/or angular momenta ($q^2 + a^2 > m^2$), the Kerr and Kerr–Newman solutions are no longer BHs but describe the gravitational fields of naked singularities. It is of interest that if we juxtapose these parameters with masses, spins and charges of known elementary particles, they correspond to such naked singularities. Surprisingly, it turns out that the above solutions with rotation do really describe the particle properties fairly well, even including some of their quantum features, see [92, 93] and references therein.

4.4 Black hole thermodynamics

4.4.1 Four laws of BH thermodynamics

By definition, a BH can emit nothing (in the framework of the classical theory) because neither photons nor massive particles are able to leave it. Nevertheless, as found for the first time by Bekenstein [29], BHs satisfy some relations identical to the well-known relations of thermodynamics, and we can ascribe to them such thermodynamic characteristics as temperature (which is in general not zero) and entropy.

Hawking [197] discovered, while considering quantum processes near horizons, that BHs emit radiation like black bodies with a certain temperature, and he derived an exact relationship between the BH entropy and the BH horizon area. A little later, four laws of BH thermodynamics were formulated [172]. As a result, there emerged a consistent description of thermodynamic systems which include both conventional matter and BHs.

The so-called *zeroth law of BH thermodynamics* is formulated as follows:

The Zeroth Law. *The surface gravity of a BH does not change from point to point of the event horizon.*

The surface gravity is, by definition, the value of the free-fall acceleration, calculated at the event horizon in terms of the particle acceleration. For a Schwarzschild BH, the surface gravity $\varkappa$ is

$$\varkappa = 1/(4m), \tag{4.91}$$

while for a Kerr BH a rather bulky calculation gives (see, e.g., [185])

$$\varkappa = \frac{r_+ - m}{2mr_+} = \frac{\sqrt{m^2 - a^2}}{2m\big(m + \sqrt{m^2 - a^2}\big)}. \tag{4.92}$$

Eqs. (4.91) and (4.92) contain only the BH parameters, without any angular dependence, thus confirming the Zeroth Law.

The First Law of BH thermodynamics reads

$$dm = \frac{\varkappa}{8\pi} dA + \Omega\, dJ, \tag{4.93}$$

where A is the horizon area,

$$A = 4\pi(r_+^2 + a^2) \tag{4.94}$$

(at $a = 0$ we obtain the expression for the Schwarzschild BH, $A = 4\pi r_+^2 = 16\pi m^2$), and Ω is the angular velocity at the horizon,

$$\Omega = \frac{a}{r_+^2 + a^2}. \tag{4.95}$$

The equality (4.93) is similar to the first law of conventional thermodynamics $dU = TdS + dW$ (U is the internal energy, S the entropy, W the work). Therefore, Eq. (4.93) may be pronounced as follows: $\Omega\, dJ$ is the work executed over the BH by adding the angular momentum dJ. Further reasoning leads to the following identifications for the BH temperature and entropy [197]:

$$T_{\rm BH} = \varkappa/(2\pi), \quad S_{\rm BH} = A/4. \tag{4.96}$$

In particular, from the expression (4.92) it follows that the temperature is zero only for an extremal BH with a double horizon ($|a| = m$).

The Second Law of BH thermodynamics is similar to the statement that the entropy cannot decrease:

The Second Law. *There is no physical process that could decrease the horizon area of a BH.*

The quantum process of BH evaporation due to Hawking radiation violates this classical law.

Lastly, the Third Law asserts the nonaccessibility of an absolute zero of temperature, or, as is sometimes formulated, the nonexistence of negative temperatures. For a BH, a similar statement reads [24]:

The Third Law. *There is no procedure able to bring the BH temperature to zero by a finite sequence of operations.*

In other words, it is impossible to make an extremal BH from a non-extremal one, to say nothing of exceeding the extremal limit (to reach $|a| > m$) and to obtain a naked singularity from a BH.

4.4.2 Black hole evaporation

As already mentioned, according to [197], BHs evaporate, emitting like black bodies with the temperature $T_{\rm BH} = \varkappa/(2\pi)$, where the surface gravity at the horizon $\varkappa$ is determined by Eqs. (4.91) and (4.92). A qualitative physical picture of BH evaporation consists in that virtual particle-antiparticle pairs, permanently appearing and disappearing according to quantum field theory, are subject to a sufficiently strong gravitational field in the neighborhood of the horizon. Then, with a certain probability (which is larger for larger space-time curvatures), the gravitational field can drag apart such virtual pairs, so that one of the particles falls to the horizon while the other escapes to infinity. As particles with nonzero energy leave the BH, the latter loses its mass.

A substitution of the corresponding constants leads to the following approximate expression for the temperature of a Schwarzschild BH in kelvins:

$$T_{\rm BH} \approx 10^{-7} m_\odot/m \ {\rm K}, \tag{4.97}$$

where $m_\odot$ is the solar mass. Thus the temperatures of BHs of solar (*a fortiori*, galactic) masses is very small, and their evaporation occurs very slowly: in fact, they increase their masses much more rapidly due to accretion of ambient matter. The smaller a BH, the larger is the curvature of space near its horizon and higher its temperature. In particular, the temperature becomes smaller if mass is added, thus BHs have a negative heat capacity.

The loss of mass (energy) of a BH can be calculated according to the Stefan-Boltzmann law, $-\dot{E}/A = \sigma T_{\rm BH}^4$, where σ is the Stefan-Boltzmann constant. Integration of this expression leads to the following BH mass as a function of time:

$$m(t) = (m_0^3 - 3Kt)^{1/3}, \quad K = \frac{1}{15360\pi} = \frac{\hbar c^4}{15360\pi G^2}, \tag{4.98}$$

where m_0 is the BH mass at $t = 0$; after the last equality sign, the expression is given in conventional units. At the time instant $t_f = m_0^3/(3K)$, the evaporation terminates, but it is so far not clear whether the BH must completely disappear or if there remains a certain object with a mass of Planckian order of magnitude, $m \sim 10^{-5}$ g. The point is that as the mass approaches its Planck value (hence also the horizon radius approaches the Planck length $\sim 10^{-33}$ cm), the so far unknown law of quantum gravity must come into force.

A second before the end of the process, as one can easily find, the BH mass is of the order 10^9 g $\sim 10^{33}$ GeV. It is the energy released in the last second of BH evaporation.

On the other hand, if one equates the lifetime $t_f = m_0^3/3K$ of a BH with a given mass m_0 to the age of the Universe, $t_U \sim 13 \cdot 10^9$ years, we obtain the mass $m_0 = m_U \sim 10^{15}$ g. Thus all primordial BHs of mass smaller than 10^{15} g, having formed in the first seconds (or millennia) of the Universe existence, have already evaporated and contributed to the chemical composition of the modern Universe, while the remaining BHs of masses close to m_U must now and then explode in the modern epoch. Heavier BHs will have a much longer life.

4.5 Regular black holes and black universes

4.5.1 Different kinds of regular black holes

One of the long-standing problems in black hole (BH) physics, and GR as a whole, is the existence of curvature singularities beyond the event horizon in the BH solutions obtained under the simplest and the most natural physical conditions (the Schwarzschild, Reissner–Nordström, Kerr and Kerr-Newman solutions of general relativity and their counterparts in many alternative theories of gravity). Singularities are places where general relativity (or another classical theory of gravity) does not work since they do not belong to a Riemannian space-time. Therefore, a full understanding of BH physics requires avoidance of singularities or/and modification of the corresponding classical theory and addressing quantum effects. There have been numerous attempts on this trend, many of them being connected with the hope that the singularities of classical gravity will be cured by effects of quantum gravity. However, of great interest also are the opportunities of singularity avoidance in the framework of classical gravity.

Let us discuss the possible geometry of classical nonsingular BHs, restricting ourselves to asymptotically flat static, spherically symmetric configurations. We begin with the general static, spherically symmetric metric

$$ds^2 = A(\rho)dt^2 - \frac{d\rho^2}{A(\rho)} - r^2(\rho)d\Omega^2, \tag{4.99}$$

written in terms of the quasiglobal coordinate ρ, which is, as we have already seen, particularly convenient for dealing with Killing horizons. At a flat asymptotic ($\rho \to \infty$) we have, without loss of generality, $A(\rho) \to 1$ and $r(\rho) \approx \rho$. A centre $\rho = \rho_c$ (if any) corresponds to $r = 0$ in a static region; horizons (if any) are described by regular zeros of the function $A(\rho)$, and their number, order and disposition determine the global causal structure of space-time.

The four simplest types of regular 4-dimensional, asymptotically flat, and static, and spherically symmetric BHs known in the literature can be presented as follows (see Fig. 4.4):

1. BHs with a regular centre ($r \approx \text{const}\,\rho - \rho_c$, $A \to A_c > 0$, $A(dr/d\rho)^2 \approx 1 + O(r^2)$ as $\rho \to \rho_c$). Since a regular centre can only be located in a R-region, such a BH must have at least two simple horizons or one double horizon, and its causal structure is then represented by the same Carter–Penrose diagram as that of the nonextreme or extreme Reissner-Nordström BH (diagrams 1b and 1c in Fig. 4.4), respectively. A larger number of horizons is not excluded, leading to more complex causal structures.

2. BHs without a centre, having second-order horizons of infinite area (so-called cold BHs because such horizons are always characterized by zero Hawking temperature) [58, 59, 97]. They may have different causal structures. In one case, the Carter–Penrose diagram coincides with that of Kerr's extreme BH, consisting of an infinite tower of R-regions but

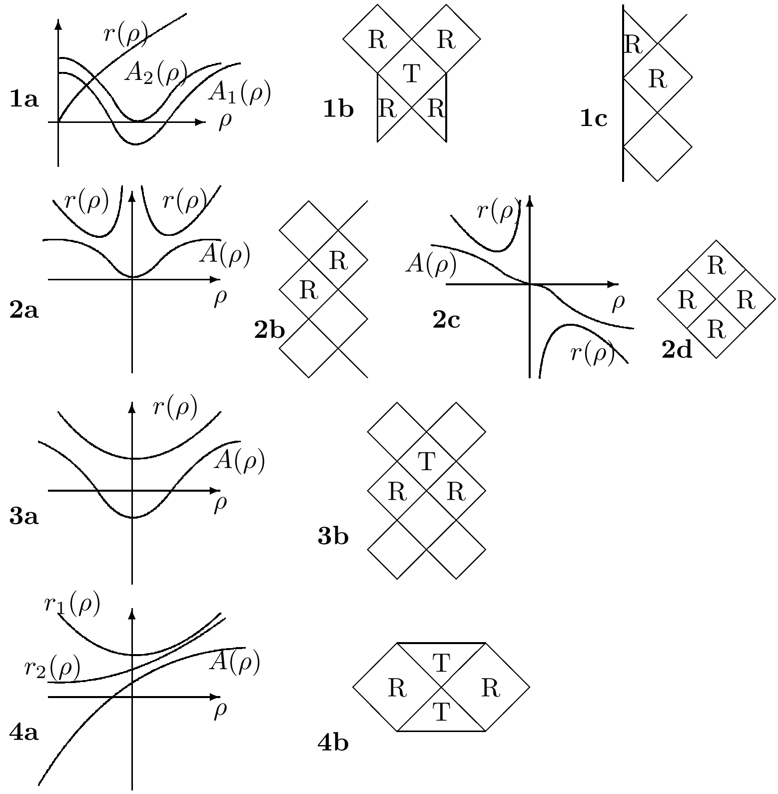


Figure 4.4 Plots showing the qualitative behaviour of the metric functions and Carter-Penrose diagrams for four different types of regular static, spherically symmetric black holes. Diagrams 1b and 1d (like those for the nonextreme and extreme Reissner–Nordström metrics) refer to curves A_1 and A_2 in plot 1a, respectively. Diagram 2b, like that for the extreme Kerr metric, refers to plot 2a, diagram 2d to plot 2c, 3b to 3a and 4b to 4a. The R and T letters in the diagrams designate the R and T space-time regions. Diagrams 1b, 1c, 2b and 3b are infinitely extendible upward and downward. In all diagrams, all inner slanting lines depict horizons while all boundaries correspond to $r = \infty$, with the following exceptions: the verticals in diagrams 1b and 1d describe a regular centre, $r = 0$; the horizontal lines in diagram 4b correspond either to $r = \infty$ or to $r = r_0 > 0$, according to the curves $r_1(\rho)$ or $r_2(\rho)$ at large negative ρ.

certainly without a ring singularity which is present in Kerr's solution (plots 2a and diagram 2b in Fig. 4.4). In another case, there are only four R-regions (plots 2c and diagram 2d).

3. BHs whose causal structure coincides with that of a nonextreme Kerr BH, again without a singular ring [60, 104] (diagram 3b).

4. Regular BHs with a Schwarzschild-like causal structure [66] (diagram 4b) but with cosmological expansion instead of a singularity. Such BHs have been termed ***black universes*** [61].

Type 1 of regular BHs traces back to Bardeen's work [23] which puts forward the very idea of regular BHs instead of singular ones and suggested, as an example, a particular BH configuration with

$$r \equiv \rho, \quad A(\rho) = 1 - \frac{M\rho^2}{(\rho^2 + q^2)^{3/2}}, \tag{4.100}$$

where $M,\ q = \text{const}$, and two horizons exist provided $q^2 < (16/27)M^2$. Later on there appeared numerous examples ([17, 52, 151] and others) of regular BH solutions where, just as in Eq. (4.100), $r \equiv \rho$. This, by virtue of the Einstein equation, implies that the stress-energy tensor of the matter source satisfies the condition

$$T^0_0 \equiv T^1_1 \tag{4.101}$$

or, in other words, $\varepsilon = -p_r$ ($\varepsilon = T^0_0$ is the energy density and $p_r = -T^1_1$ is the radial pressure). The condition (4.101) is invariant under radial boosts, making it possible to ascribe the source to vacuum matter [151]. Indeed, it is a definitive property of vacuum that there is no comoving reference frame: this kind of matter looks the same whatever is our motion in the radial direction. At the regular centre in this case the matter equation of state has necessarily the form of a cosmological constant [62], $T^\nu_\mu \propto \delta^\nu_\mu$.

It has also been shown [52] that regular BHs with $r \equiv \rho$ and any $A(\rho)$ satisfying the regular centre conditions may be obtained as magnetic monopole solutions of general relativity coupled to gauge-invariant nonlinear electrodynamics with the Lagrangian $L(F)$, $F := F_{\mu\nu}F^{\mu\nu}$ ($F_{\mu\nu}$ is the electromagnetic field tensor): the arbitrariness in $A(\rho)$ corresponds to the freedom of choosing the function $L(F)$. Solutions with an electric charge were shown [52] to be impossible whatever the choice of $L(F)$ if $L(F)$ is the same in the whole space; this theorem, however, may be circumvented by assuming different forms of $L(F)$ near the centre and at large r [94], i.e., by requiring a sort of phase transition(s) at some value(s) of the radial coordinate.

Other examples of type 1 regular BHs have also been found and discussed, see, e.g., [16, 33, 200, 268, 291, 300, 368, 378] and references therein. Among them, the BH models [16, 300, 378] are related to different ideas of quantum gravity (the so-called UV-complete quantum gravity and noncommutative geometry) realized in a semiclassical approximation.

Type 2 regular BHs, with and without an electric charge, have been obtained [58, 59, 97] in the framework of the Brans-Dicke scalar-tensor theory (STT) with the coupling constant $\omega < -3/2$, when the theory is of anomalous, or phantom nature. Their existence requires fine tuning in the form of specific relations between ω and the integration constants of the corresponding exact solutions. It is of interest to note that these regular cold Brans–Dicke BHs have singular counterparts in the Einstein frame, in other words, in general relativity with a minimally coupled phantom scalar field (see the anti-Fisher solution (4.41)) [57].

Type 3 regular BHs have been found [60, 104] as static, spherically symmetric solutions to the effective equations [371] describing 4D gravity in an RS2 type brane world. It has been shown that such regular solutions are generic in a certain range of the integration constants [60] and that many of them are stable, at least under some kinds of perturbations [2]. It should be stressed, however, that these 4D equations do not form a closed set; to study the full 5D geometry of the bulk one should solve the corresponding 5D equations, and there are only tentative results in this direction [105].

Type 4 configurations have been obtained [66] as generic solutions to the Einstein-scalar equations for the case of minimally coupled phantom scalar fields with certain potentials. We would remind the reader that such scalar fields have recently become popular in the cosmological context since they are able to provide an equation of state with the pressure to density ratio $p/\varepsilon = w < -1$. It is this kind of equation of state that is probably required for the dark energy (DE) component of the material content of the Universe to account for its accelerated expansion (see, e.g., the reviews [120, 363] and references therein). Type 4 configurations, termed black universes [61], are, in our view, of particular interest, and in what follows we will discuss them in some more detail.

Let us note that the above list of structures is certainly incomplete: thus, one could easily imagine regular configurations with a larger number of horizons (e.g., draw in a proper place a smooth peak twice crossing the ρ axis in any plot of $A(\rho)$ in Fig. 4.4).

An interesting example of a structure somewhat similar to a black universe has recently appeared in studies on the basis of a semiclassical limit of loop quantum gravity [299, 323], see Fig. 4.5. This "quantum-corrected Schwarzschild black hole" also has a Kantowski–Sachs temporal infinity instead of the singularity $r = 0$, but there are at least three important points in which it differs from our black universes: (1) the minimum radius

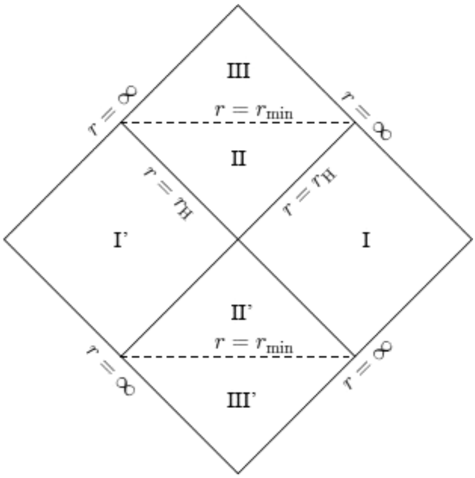


Figure 4.5 A regular BH structure obtained in [323]. I and I' are R-regions like those in the Schwarzschild space-time, II and III are Kantowski–Sachs regions (II' and III expanding, II and III' contracting).

$r_{\min}$ is always reached in the T-region; (2) $r_{\min}$ is of the order of the Planck length whereas in our models it is arbitrary, and (3) this configuration does not become isotropic at late times like ours.

Other conclusions on a quantum-corrected Schwarzschild space-time have been obtained in [42] (approaching at large times a Nariai-type metric with a constant spherical radius) and [16, 300, 378] (acquiring a regular centre). This diversity is one of manifestations of an as yet uncertain choice of a correct theory of quantum gravity.

4.5.2 Black universes with a minimally coupled scalar field

Consider the action for a self-gravitating phantom scalar field in general relativity

$$S = \int \sqrt{g}\, d^4x [R - (\partial\phi)^2 - 2V(\phi)], \tag{4.102}$$

where $g = |\det(g_{\mu\nu})|$, $(d\phi)^2 = g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi$ and $V(\phi)$ is an arbitrary potential. With the metric (4.99) and $\phi = \phi(\rho)$, the scalar field equation and three independent combinations of the Einstein equations read

$$(Ar^2\phi')' = -r^2 dV/d\phi, \tag{4.103}$$

$$(A'r^2)' = -2r^2 V, \tag{4.104}$$

$$2r''/r = \phi'^2, \tag{4.105}$$

$$A(r^2)'' - r^2 A'' = 2, \tag{4.106}$$

where the prime denotes $d/d\rho$. The scalar field equation (4.103) is a consequence of Eqs. (4.104)–(4.106), which, given a potential $V(\phi)$, form a determined set of equations for the unknowns $r(\rho)$, $A(\rho)$, $\phi(\rho)$. Eq. (4.106) can be integrated, giving

$$B' \equiv (A/r^2)' = 2(\rho_0 - \rho)/r^4, \tag{4.107}$$

where $B(\rho) = A/r^2$ and ρ_0 is an integration constant.

As we know from Sec. 4.2.2, Eq. (4.107) severely restricts the possible dispositions of zeros of the function $A(\rho)$ and hence the global causal structure of space-time [53], leading to the Global Structure Theorem (see Sec. 4.3), according to which the choice of possible types of global causal structure is precisely the same as for the general Schwarzschild-de Sitter solution with arbitrary mass and cosmological constant. This result equally applies to normal and phantom fields since Eqs. (4.106) and (4.107) are the same for them. It holds for any sign and shape of $V(\phi)$ and under any assumptions on the asymptotics. BHs with scalar hair (respecting the existing no-hair theorems) are not excluded. Examples of (singular) BHs with both normal (e.g., [84, 108, 316]) and phantom [126] scalar hair are known. However, BHs with a regular centre (type 1 according to the previous section) are ruled out for our system since their existence requires a regular minimum of $B(\rho)$.

As shown in Ref. [66], the system (4.103)–(4.106) has as many as 16 types of regular solutions with flat, de Sitter and AdS asymptotic behaviours. Let us discuss asymptotically flat configurations, for which $A(\rho) \to 1$ and $r(\rho) \approx \rho$ as $\rho \to \infty$. Then, as ρ decreases from infinity, the derivative r' also decreases according to Eq. (4.105), the decrease rate depending on the details of the system. If the decrease is slow enough, $r(\rho)$ will reach zero at some finite ρ. It can be a regular centre only if horizons are absent (otherwise there would be a type 1 regular BH which is ruled out here), and a particle-like solution is then obtained instead of a BH one.

Assuming the absence of singularities, other opportunities are $r(\rho) \to r_0 = \text{const}$ and $r(\rho) \to \infty$ as $\rho \to -\infty$. (Note that any kind of oscillatory behaviour of $r(\rho)$ is ruled out by (4.105) according to which $r'' \geq 0$.)

In the first case, according to (4.107),

$$A' \approx -2\rho/r_0^2 \to +\infty, \quad A \approx -\rho^2/r_0^2 \quad \text{as } \rho \to -\infty.$$

So this "r_0 asymptotic" is located in a T-region. The radius r_0 is related to the limiting value of the potential: $V \to -1/r_0^2$. Changing the notations $-\rho \to T$ and $t \to x$ (since ρ here becomes a temporal coordinate and the former time t a spatial one), we can write the asymptotic form of the resulting Kantowski–Sachs metric as

$$ds^2 \approx \frac{r_0^2}{T^2}\, dT^2 - \frac{T^2}{r_0^2} dx^2 - r_0^2 d\Omega^2 \quad \text{as } T \to \infty. \tag{4.108}$$

It is a highly anisotropic universe, exhibiting no expansion in the two angular directions and an exponential (in terms of the physical time $\tau \sim \log T$) expansion in the third direction x. From the viewpoint of an observer at large positive ρ, this universe is located beyond the event horizon of a BH.

In the case $r \to \infty$, the most interesting opportunity is that $r \sim |\rho|$ as $\rho \to -\infty$. Assuming $r \approx -a\rho$, $a = \text{const} > 0$, we obtain from (4.107) and (4.104):

$$A \approx 1/a^2 - Ca^2\rho^2, \quad V \approx 3Ca^2, \quad C = \text{const}. \tag{4.109}$$

It is easy to verify that $C = 0$ leads to a Minkowski metric at large negative ρ (though the time rate will be different from that at large positive ρ if $a \neq 1$), an anti-de Sitter metric if $C < 0$ and a de Sitter metric if $C > 0$. In the cases $C \leq 0$ horizons are absent since otherwise the function $B(\rho)$ would have a minimum at some finite ρ, which cannot happen due to (4.107). Thus possible solutions with $C \leq 0$ describe traversable wormholes (see Chapter 5), and the details of their geometry depend on the particular choice of $V(\phi)$.

Lastly, for $C > 0$ we obtain a de Sitter asymptotic behavior of the solution, describing (since we are now in a T-region) isotropic expansion or contraction. From the viewpoint of an external observer, located at large positive ρ, it is a BH, but a possible BH explorer now has a chance to survive for a new life in an expanding, gradually isotropizing Kantowski–Sachs universe. The specific BH profile and the isotropization regime after crossing the horizon depend on the choice of $V(\phi)$.

Since, according to the Global Structure Theorem (see above), an asymptotically flat configuration can have only one simple horizon, such a BH has a Schwarzschild-like causal structure, but the singularity $r = 0$ in the Carter-Penrose diagram is now replaced by the de Sitter infinity $r = \infty$.

A simple example may be obtained by putting [66]

$$r = (\rho^2 + b^2)^{1/2}, \quad b = \text{const} > 0, \tag{4.110}$$

and using the inverse problem scheme. Eq. (4.107) gives

$$B(\rho) = \frac{A(\rho)}{r^2(\rho)} = \frac{c}{b^2} + \frac{1}{b^2 + \rho^2} + \frac{\rho_0}{b^3}\left(\frac{b\rho}{b^2+\rho^2} + \tan^{-1}\frac{\rho}{b}\right), \tag{4.111}$$

where $c = \text{const}$. Eqs. (4.105) and (4.104) then lead to expressions for $\phi(\rho)$ and $V(\rho)$:

$$\phi = \pm\sqrt{2}\tan^{-1}(\rho/b) + \phi_0, \tag{4.112}$$

$$V = -\frac{c}{b^2}\frac{r^2 + 2\rho^2}{r^2} - \frac{\rho_0}{b^3}\left(\frac{3b\rho}{r^2} + \frac{r^2+2\rho^2}{r^2}\tan^{-1}\frac{\rho}{b}\right), \tag{4.113}$$

with $r = r(\rho)$ given by (4.110). In particular,

$$B(\pm\infty) = -\frac{1}{3}V(\pm\infty) = \frac{2bc \pm \pi\rho_0}{2b^3}. \tag{4.114}$$

Choosing in (4.112), without loss of generality, the plus sign and $\phi_0 = 0$, we obtain for $V(\phi)$ ($\psi := \phi/\sqrt{2}$):

$$V(\phi) = -\frac{c}{b^2}(3 - 2\cos^2\psi) - \frac{\rho_0}{b^3}[3\sin\psi\cos\psi + \psi(3 - 2\cos^2\psi)]. \tag{4.115}$$

The solution behaviour is controlled by two integration constants: c that moves the plot of $B(\rho)$ up and down, and ρ_0 showing the maximum of $B(\rho)$. Both $r(\rho)$ and $B(\rho)$ are even functions if $\rho_0 = 0$, otherwise $B(\rho)$ loses this symmetry. Asymptotic flatness at $\rho = +\infty$ implies $2bc = -\pi\rho_0$ while the Schwarzschild mass, defined in the usual way, is $m = \rho_0/3$.

Under this asymptotic flatness assumption, for $\rho_0 = m = 0$ we obtain the simplest symmetric configuration, the Ellis wormhole [155]: $A \equiv 1$, $V \equiv 0$. For $\rho_0 < 0$, according to (4.114), we obtain a wormhole with $m < 0$ and an AdS metric at the far end, corresponding to the cosmological constant $V_- < 0$. For $\rho_0 > 0$, when $V_- > 0$, there is a regular BH with $m > 0$ and a de Sitter asymptotic far beyond the horizon, precisely corresponding to the above description of a black universe.

The horizon radius $r(\rho_h)$ can be obtained by solving the transcendental equation $A(\rho_h) = 0$, where $A(\rho)$ is given by Eq. (4.111). It depends on the parameters m and $b = \min r(\rho)$ and cannot be smaller than b, which plays the role of a scalar charge: $\psi \approx \pi/2 - b/\rho$ at large ρ. Since $A(0) = 1 + c$, the throat $\rho = 0$ is situated in the R-region if $c > -1$, i.e., if $3\pi m < 2b$, precisely at the horizon if $3\pi m = 2b$ and in the T-region beyond the horizon if $3\pi m > 2b$. These relations between the constants m and b prompt (and this is probably the case in more general situations) that if the BH mass dominates over the scalar charge, there is no throat in the static region, and for a distant observer the BH looks approximately the same as usual in general relativity.

As follows from Eqs. (4.103) and (4.104), the potential V tends to a constant and, moreover, $dV/d\phi \to 0$ at each end of the ρ range. It is a general property of all classes of regular solutions indicated in [66]. More precisely, *a regular scalar field configuration requires a potential with at least two zero-slope points (not necessarily extrema) at different values of* ϕ.

Suitable potentials are, e.g., $V = V_0 \cos^2(\phi/\phi_0)$ and the Mexican hat potential $V = (\lambda/4)(\phi^2 - \eta^2)^2$ where V_0, ϕ_0, λ, η are constants. A flat infinity at $\rho = +\infty$ certainly requires $V_+ = 0$, while a de Sitter asymptotic can correspond to a maximum of V since phantom fields tend to climb up the slope of the potential rather than roll down, as is evident from Eq. (4.103). Accordingly, Faraoni [160], considering spatially flat isotropic phantom cosmologies, has shown that if $V(\phi)$ is bounded above by $V_0 =$ const > 0, the de Sitter solution is a global attractor. Very probably this conclusion extends to Kantowski–Sachs cosmologies after isotropization.

One more point may be stressed: the late-time de Sitter expansion rate is entirely determined by the corresponding potential value $V_- > 0$ (which, in our notation according to (4.102), coincides with the effective cosmological constant Λ at late times) rather than by the details of the solution such as the Schwarzschild mass defined at the flat asymptotic.

We can conclude that black universes are a generic kind of solutions to the Einstein-scalar equations in the case of phantom scalars with proper potentials.

Thus we have classified the possible geometries of static, spherically symmetric, and asymptotically flat BHs and discussed in more detail a particular type, the black universes. Such hypothetical configurations combine the properties of a wormhole (absence of a centre, a regular minimum of the area function) and a black hole (a Killing horizon separating R- and T-regions).

Quite naturally, such unusual objects require unusual matter for their existence. We saw how they can be obtained using a phantom scalar field. As shown in [61], such solutions also exist in much more general frameworks for the description of phantom matter, namely, STT of gravity and the so-called k-essence. A conclusion is that a great number of models of phantom matter produce such solutions.

The latter lead to the idea that our Universe could appear from phantom-dominated collapse in another, "mother" universe and undergo isotropization (e.g., due to particle creation) soon after crossing the horizon. It is known that a Kantowski–Sachs nature of our Universe is not excluded observationally [8] if its isotropization had happened early enough, before the last scattering epoch (at redshifts $z \gtrsim 1000$). One can notice that we are thus facing one more mechanism of universes multiplication, in addition to the well-known mechanism existing in the chaotic inflation scenario.

But, as we saw in Sec. 4.1.1, there are arguments both *pro et contra* phantom fields, and the latter seem somewhat stronger. It is reasonable to try to avoid such fields in modelling real or hypothetic objects and phenomena. In what follows, we describe two such attempts: one [63] is devoted to obtaining a black universe model in a brane-world scenarios of RS2 type; another [64] uses the notion of a "trapped ghost" [88], i.e., a scalar field which has phantom properties only in a restricted region of space, a strong-field region, whereas far away from it all standard energy conditions are observed. This can explain why ghosts, or phantoms, are not observed under usual physical conditions but certainly does not remove the basic shortcomings of phantom fields, see Sec. 4.1.1.

4.5.3 A black universe in a brane world

The brane world concept describes our world as a 4D surface (brane) supporting all or almost all matter fields and embedded in a higher-dimensional space-time (called the bulk). This concept traces back to the early 80s and leads to a variety of consequences in cosmology, gravitational and particle physics, see the reviews [119, 283, 311, 350]. In particular, brane worlds turn out to be a natural framework for wormholes (see Sec. 5.3.3) since there the modified Einstein equations [371] (4.116) contain a source term E_μ^ν of geometric origin which need not observe the usual energy conditions. As we shall see now, it is this source term that can replace phantom fields in constructing black-universe models.

The modified Einstein equations (4.116) used here correspond to the so-called RS2 scenario: a single brane in a $\mathbb{Z}_2$-symmetric 5-dimensional bulk, with all fields except gravity confined to the brane. It generalizes the second Randall–Sundrum model comprising a single Minkowski brane in an AdS bulk [337]. In other brane-world scenarios, the effective 4D Einstein equations also contain terms similar to E_μ^ν, e.g., on codimension-1 branes without Z_2 symmetry [431] and in Dvali–Gabadadze–Porrati brane worlds [150] with different kinds of induced gravity terms [360]. Thus we can anticipate that black-universe solutions similar to ours exist in such brane worlds as well, though probably under some other conditions.

The gravitational field on the brane is described by the modified Einstein equations [371]

$$G_\mu^\nu = -\Lambda_4 \delta_\mu^\nu - \varkappa_4^2 T_\mu^\nu - \varkappa_5^4 \Pi_\mu^\nu - E_\mu^\nu, \tag{4.116}$$

where $G_\mu^\nu = R_\mu^\nu - \frac{1}{2}\delta_\mu^\nu R$ is the 4D Einstein tensor, Λ_4 is the 4D cosmological constant expressed in terms of Λ_5 and the brane tension λ:

$$\Lambda_4 = \frac{1}{2}\left(\Lambda_5 + \frac{1}{6}\varkappa_5^4 \lambda^2\right); \tag{4.117}$$

$\varkappa_4^2 = 8\pi G_N = \varkappa_5^4 \lambda/(6\pi) = m_4^{-2}$ is the 4D gravitational constant; G_N is Newton's constant of gravity, and m_4 is the 4D Planck mass;

$\varkappa_5 = m_5^{-3/2}$, m_5 being the 5D Planck energy scale;

T_μ^ν is the SET of matter trapped on the brane;

Π_μ^ν is a tensor quadratic in T_μ^ν, obtained from matching the 5D metric across the brane:

$$\Pi_\mu^\nu = \frac{1}{4} T_\mu^\alpha T_\alpha^\nu - \frac{1}{2} T T_\mu^\nu - \frac{1}{8}\delta_\mu^\nu \left(T_{\alpha\beta}T^{\alpha\beta} - \frac{1}{3}T^2\right), \tag{4.118}$$

where $T = T_\alpha^\alpha$; and
E_μ^ν is the "electric" part of the 5D Weyl tensor projected onto the brane: in proper 5D coordinates,

$$E_{\mu\nu} = \delta_\mu^A \delta_\nu^C {}^{(5)}C_{ABCD} n^B n^D, \tag{4.119}$$

where $A, B, \ldots$ are 5D indices and n^A is the unit normal to the brane. By construction, E_μ^ν is traceless, and $E_\mu^\mu = 0$.

Other characteristics of E_μ^ν are unknown without specifying the 5D metric, hence the set of equations (4.116) is not closed. In isotropic cosmology this leads to an additional arbitrary constant in the field equations, connected with the density of "dark radiation". For static, spherically symmetric systems to be discussed here, this freedom is expressed in a single arbitrary function of the radial coordinate.

Reasons for neglecting Π_μ^ν

Let us show that under quite reasonable conditions we can neglect the tensor Π_μ^ν in (4.116).

We put $\Lambda_4 = 0$, so that

$$|\Lambda_5| = \frac{1}{6}\varkappa_5^4\lambda^2 = 6\pi^2(\varkappa_4/\varkappa_5)^4, \qquad (4.120)$$

and use the observational restriction on the bulk length scale ℓ which follows from the recent short-range Newtonian gravity tests [279], showing that Newton's inverse-square law holds at length scales greater than about 0.1 mm. This means that if we live on an RS2-like brane, the bulk length scale can be estimated as

$$\ell = (6/|\Lambda_5|)^{1/2} \lesssim 10^{-2} \text{ cm}. \qquad (4.121)$$

Note that the 4D Planck scale in our notation is

$$m_4 = \varkappa_4^{-1} = 8\pi G_\text{N} \approx 2.4{\cdot}10^{18}\,\text{GeV},$$
$$l_4 = 1/m_4 = \varkappa_4 \approx 8{\cdot}10^{-33}\,\text{cm}.$$

Combining (4.120) and (4.121), we obtain

$$m_5/m_4 = (\pi\ell/l_4)^{-1/3} \gtrsim 10^{-10}, \qquad (4.122)$$

so that the 5D Planck energy scale in this scenario is at least about 10^8 GeV.

Now, we can assert that the term with Π_μ^ν is negligible in (4.116) as compared with the T_μ^ν term as long as

$$\varkappa_5^4 W^2 \ll \varkappa_4^2 W,$$

where W characterizes the magnitude of T_μ^ν, say, the absolute value of the largest component of T_μ^ν or

$$W \ll m_5^6/m_4^2 = m_4^4(m_5/m_4)^6, \qquad (4.123)$$

where $m_4^4 \approx 3.5{\cdot}10^{73}\,\text{GeV}^4 \approx 8.4{\cdot}10^{90}\ \text{g}{\cdot}\text{cm}^{-3}$ is the Planck density while the second factor is, according to the experimental bound (4.122), about 10^{-60} or larger. As a result, for the "density" W we have

$$W \ll 10^{30}\ \text{g}{\cdot}\text{cm}^{-3}.$$

Recalling that the density of nuclear matter is about $10^{13}\ \text{g}{\cdot}\text{cm}^{-3}$, it is clear that this bound certainly holds for any thinkable matter.

Brane gravity with a scalar field

Consider Eq. (4.116) for static, spherically symmetric configurations of a normal scalar field with an arbitrary potential, neglecting the term Π_μ^ν. So the scalar field Lagrangian has the form (4.47). The general static, spherically symmetric 4D metric is again taken in the form

$$ds^2 = A(u)dt^2 - du^2/A(u) - r^2(u)d\Omega^2 \tag{4.124}$$

with the quasiglobal coordinate u.

The scalar field SET is conservative, so the same is required for E_μ^ν. If we take it, for convenience, in the form

$$E_\mu^\nu = \mathrm{diag}(-P{-}Af,\ -P,\ P{+}Af/2,\ P{+}Af/2), \tag{4.125}$$

where P and f are some functions of the radial coordinate u (so that, as required, its trace is zero), then the conservation law $\nabla_\alpha E_1^\alpha$ is written as

$$(Pr^4)' = \frac{1}{2}r^6 f(A/r^2)'. \tag{4.126}$$

Equations (4.116) may be written in the form

$$\frac{1}{2r^2}(A'r^2)' = -V - P - Af, \tag{4.127}$$

$$2r''/r = -\phi'^2 + f, \tag{4.128}$$

$$A(r^2)'' - r^2A'' - 2 = 2P + \frac{3}{2}Af. \tag{4.129}$$

The scalar field equation $(Ar^2\phi')' = r^2 dV/d\phi$ follows from (4.127)–(4.129) combined with (4.126).

Thus, if $V(\phi)$ is specified, we have four independent equations (4.126)–(4.129) for five unknown functions of u: ϕ, A, r, f and P. The system becomes still more underdetermined if $V(\phi)$ is not specified: in this case we can choose as many as two functions by hand to obtain a solution.

To have a regular positive minimum of $r(u)$, required for a black universe, we must have $r'' > 0$ at least in some range of u. By (4.128), this can be achieved only with $f > 0$. However, if we try to put $r'' > 0$ *everywhere*, i.e., $f > \phi'^2$, it inevitably leads to an oscillatory behavior of $A(u)$, with at least three horizons instead of the desired one [63]. A way of avoiding such a behavior is to try a "more concentrated" distribution of f and P.

So we first try $f(u)$ maximally peaked on a single sphere, for example, $u = 0$, i.e., proportional to $\delta(u)$, the Dirac delta function. We now make the equations dimensionless by putting

$$\begin{aligned} &x = u/b, \quad Pr^4 = Q(x)b^2, \quad fr^4 = 2F(x)b^2, \\ &r = b\bar{r}(x), \quad B = \bar{B}(x)/b^2, \end{aligned} \tag{4.130}$$

where $b = \text{const} > 0$ specifies a length scale of the configuration, and to justify our rejection of Π_μ^ν we assume $b \gg \ell$ (see above).

Omitting the bars over r and B, we write Eqs. (4.126), (4.128) and (4.129) in the form

$$Q' = r^2 F B', \tag{4.131}$$

$$\frac{r''}{r} = -\psi'^2 + \frac{F}{r^4}, \tag{4.132}$$

$$(r^4 B')' + 2 + \frac{4Q}{r^2} + 6BF = 0, \tag{4.133}$$

where the prime denotes d/dx and $\psi = \phi/\sqrt{2}$.

Then, we choose $r(x)$ so that $r'' < 0$ at all $x \neq 0$:

$$r^2(x) = (|x| + c)^2 - 1, \quad c = \text{const} > 1. \tag{4.134}$$

This conforms to both flat and de Sitter asymptotics. Then, in Eq. (4.132), the quantity r''/r has a delta-like singularity at $x = 0$. To compensate it and make ψ' continuous at $x = 0$, we put

$$F(x) = 2cr_0^2\delta(x), \quad r_0 = \sqrt{c^2 - 1}. \tag{4.135}$$

We then have

$$\psi' = \pm\frac{1}{r^2} = \frac{1}{(|x| + c)^2 - 1}, \tag{4.136}$$

whence, without loss of generality,

$$\begin{aligned} &\psi = \begin{cases} h(c - x), & x < 0, \\ 2h(c) - h(c + x), & x > 0, \end{cases} \\ &h(x) := \frac{1}{2}\ln\left|\frac{x + 1}{x - 1}\right|. \end{aligned} \tag{4.137}$$

Equation (4.131) gives for $Q(x)$:

$$Q(x) = Q_0 + 2cr_0^4 B_{x0}\theta(x), \tag{4.138}$$

where $Q_0 = \text{const}$, $B_{x0} = B'(0)$ and $\theta(x)$ is Heaviside's function equal to zero at $x < 0$ and to unity at $x > 0$.

It remains to solve Eq. (4.133) and to find V. We have solved it choosing the initial conditions at large negative x in the Schwarzschild form

$$A(x) = r^2(x)B(x) = 1 - \frac{2m}{|x|}, \quad m = \text{const.} \tag{4.139}$$

The solution [63] is analytical but rather bulky and will not be presented here. It really describes a black universe, but the delta-like distribution of the effective exotic matter, characterized by $f(u)$, causes an undesirable discontinuity of B' at $x = 0$.

Evidently the qualitative behavior of the model will not change if we replace the delta-like distribution of $f(x)$ with a smooth one but sufficiently peaked near $x = 0$. We will present an example of such a solution. Namely, let us preserve the notations (4.130), so that the field equations have the form (4.131)–(4.133); however, instead of (4.134), we choose the following function $r(x)$:

$$r^2(x) = (|x| + 1)^2 - 1, \quad |x| > d, \tag{4.140}$$

$$r^2(x) = ax^2 + s = \frac{d+1}{d}x^2 + d, \quad |x| < d, \tag{4.141}$$

where $d = \text{const} > 0$ is sufficiently small and the constants in (4.141) are chosen to make r and r' continuous at $x = \pm d$, as shown in Fig. 4.6.

Let us take $f(x) \equiv 0$ at $|x| > d$ and, for $|x| < d$, as in (4.71), $f = 2C/r^6$. From (4.131) we find $Q(x)$ as follows:

$$Q(x) = \begin{cases} Q_0, & x < -d; \\ Q_0 + [B(x) - B(-d)]C, & |x| \le d; \\ Q_0 + [B(d) - B(-d)]C, & x > d, \end{cases} \tag{4.142}$$

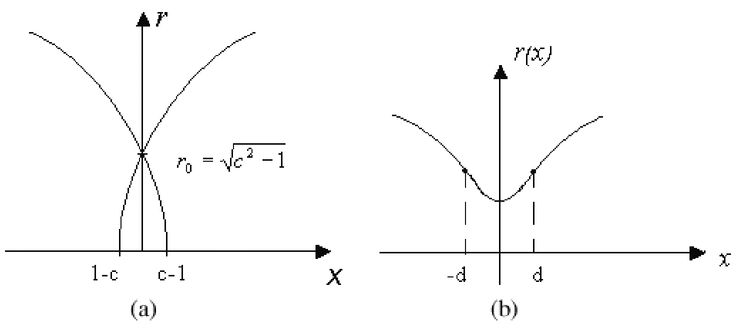


Figure 4.6 The area function $r(x)$ chosen with a delta-like (a) and smooth (b) function $f(r)$.

where Q_0 is an integration constant. The constant C (more precisely, its dimensionless counterpart $\bar{C} = C/b^4$) is determined from Eq. (4.132), where we require continuity of ψ' at $x = \pm d$. Eq. (4.133) for the function $B(x) = A/r^2$ is then solved analytically for $|x| > d$ but only numerically for $|x| < d$.

Integrating and maintaining continuity of B and B' at $x = \pm d$, we obtain $B(x)$ shown in Fig. 4.7. Figure 4.8 shows the corresponding potential $V(x)$. Clearly these are black universe models, where the asymptotic behavior of $V(x)$ approaching a positive constant as $x \to \infty$ corresponds to de Sitter expansion with a positive cosmological constant.

Thus we have built a family of black universe solutions to the modified Einstein equations valid in an RS2 type brane world without explicitly introducing any phantom matter. Instead of exotic matter in the field equations we have used the "tidal" term of geometric origin, which has no reason to respect the energy conditions known for physically plausible matter fields.

The presently obtained models do not pretend to be quite realistic; they simply show the possibility of such a scenario in principle. As any solution to the effective 4D equations describing gravity on the brane, they certainly

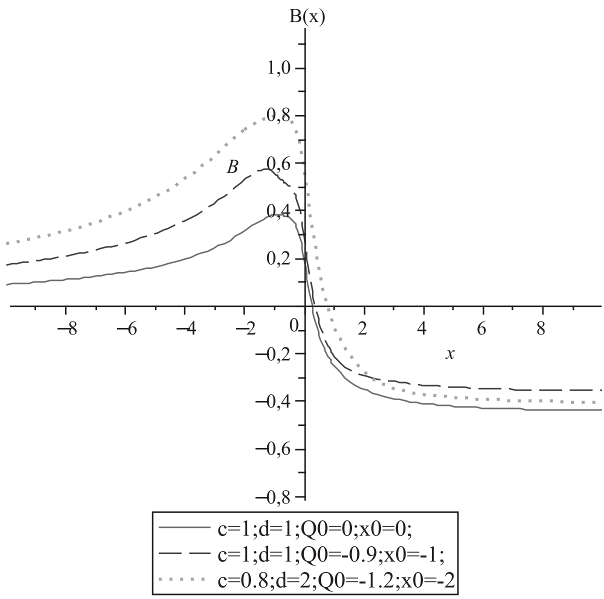


Figure 4.7 The function $B(x)$ obtained from a smoothed but peaked $f(u)$.

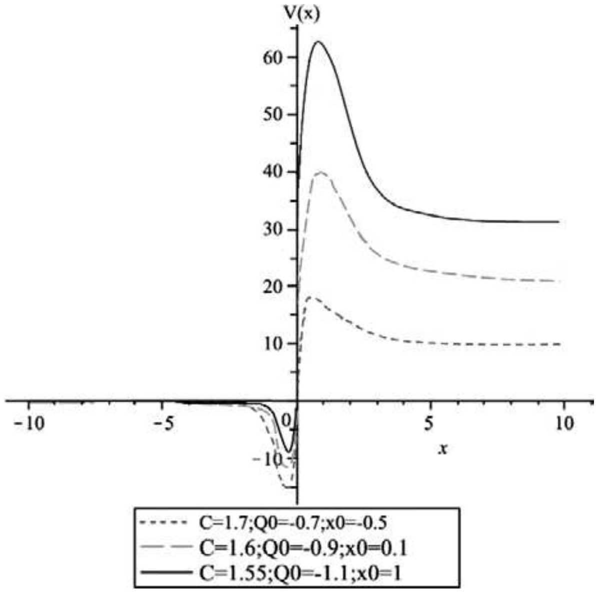


Figure 4.8 The potential $V(x)$ obtained from a smoothed but peaked $f(u)$.

need extension to the bulk, whose finding is quite a challenging problem, although an extension always exists due to the known embedding theorems. Another problem to be solved is that of stability of such solutions.

4.5.4 A black universe with a trapped ghost

Let us now discuss another opportunity [64, 88] of obtaining black universe (and wormhole) configurations in the framework of general relativity, with a kind of matter which is phantom only in a restricted strong-field region of space, whereas outside it the standard energy conditions are observed. As an example of such matter, we consider static, spherically symmetric configurations of a minimally coupled scalar field with the Lagrangian

$$L_s = \frac{1}{2}h(\phi)g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - V(\phi), \tag{4.143}$$

where $h(\phi)$ and $V(\phi)$ are arbitrary functions. If $h(\phi)$ has a variable sign, it cannot be absorbed by redefinition of ϕ in its whole range. Cases of interest are those where $h > 0$ (that is, the scalar field is canonical, with positive kinetic energy) in a weak field region and where $h < 0$ (a phantom, or ghost scalar field) in some restricted region e.g., a wormhole throat. In this sense it can be said that the ghost is trapped. A possible transition between $h > 0$ and $h < 0$ in cosmology was considered in [260].

Taking the metric in the form (4.124), let us specify which kinds of functions $r(u)$ and $A(u)$ are required for obtaining a black universe.

1. Regularity in the whole range $u \in \mathbb{R}$.
2. Asymptotic flatness as $u \to +\infty$ (without loss of generality), i.e., $r(u) \approx u$, $A(u) \to 1$;
3. A de Sitter asymptotic as $u \to -\infty$, i.e., a T-region ($A < 0$) where $r(u) \sim |u|$, $-A(u) \sim u^2$;
4. A single simple horizon (i.e., a simple zero of $A(u)$) at finite u. It is an event horizon as seen from the static side, and it is the starting point of the cosmological evolution as seen from the T-region.

The existence of two asymptotic regions $r \sim |u|$ as $u \to \pm\infty$ requires at least one regular minimum of $r(u)$ at some $u = u_0$, at which

$$r = r_0 > 0, \quad r' = 0, \quad r'' > 0, \tag{4.144}$$

where the prime stands for d/du. (In special cases where $r'' = 0$ at the minimum, we inevitably have $r'' > 0$ in its neighborhood.)

The necessity of violating the weak and null energy conditions at such minima follows from the Einstein equations. Indeed, one of them reads

$$2A\,r''/r = -(T^t_t - T^u_u), \tag{4.145}$$

where T^ν_μ are components of the stress-energy tensor (SET).

In a R-region ($A > 0$), the condition $r'' > 0$ implies $T^t_t - T^u_u < 0$; in the usual notations $T^t_t = \rho$ (density) and $-T^u_u = p_r$ (radial pressure) it is rewritten as $\rho + p_r < 0$, which manifests violation of the weak and null energy conditions. It is the simplest proof of this well-known violation near a static, spherically symmetric wormhole throat [302].

However, a minimum of $r(u)$ can occur in a T-region, and it is then not a throat but a bounce in the evolution of one of the Kantowski–Sachs scale factors (the other scale factor is $[-A(u)]^{1/2}$). Since in a T-region t is a spatial coordinate and u temporal, the meaning of the SET components is $-T^t_t = p_t$ (pressure in the t direction) and $T^u_u = \rho$; nevertheless, the condition $r'' > 0$ applied to (4.145) again leads to $\rho + p_t < 0$, violating the energy conditions. In the intermediate case where a minimum of $r(u)$ coincides with a horizon ($A = 0$), the condition $r'' > 0$ holds in its vicinity, along with all its consequences. Thus *the energy conditions are violated near a minimum of r in all cases.*

Let us now turn to the scalar field $\phi(u)$ with the Lagrangian (4.143). In a space-time with the metric (3.75) it has the SET

$$T_\mu^\nu = \frac{1}{2} h(u) A(u) \phi'(u)^2 \operatorname{diag}(1, -1, 1, 1) + \delta_\mu^\nu V(u). \tag{4.146}$$

The kinetic energy density is positive if $h(\phi) > 0$ and negative if $h(\phi) < 0$, so the solutions sought for must be obtained with $h > 0$ at large values of the spherical radius $r(u)$ and $h < 0$ at smaller radii r. It has been shown [88] that this goal cannot be achieved for a massless field ($V(\phi) \equiv 0$).

Thus we seek black universe configurations with a nonzero potential $V(\phi)$. The Einstein-scalar equations can be written as

$$(Ar^2 h\phi')' - \frac{1}{2} Ar^2 h'\phi' = r^2 dV/d\phi, \tag{4.147}$$

$$(A'r^2)' = -2r^2 V, \tag{4.148}$$

$$2r''/r = -h(\phi){\phi'}^2, \tag{4.149}$$

$$A(r^2)'' - r^2 A'' = 2, \tag{4.150}$$

$$-1 + A'rr' + Ar'^2 = r^2 \left(\frac{1}{2} hA\phi'^2 - V\right). \tag{4.151}$$

Eq. (4.147) follows from (4.148)–(4.150) which, given the potential $V(\phi)$ and the kinetic function $h(\phi)$, form a determined set of equations for the unknowns $r(u)$, $A(u)$, $\phi(u)$. Eq. (4.151) (the $\binom{1}{1}$ component of the Einstein equations), free from second-order derivatives, is a first integral of (4.147)–(4.150) and can be obtained from (4.148)–(4.150) by excluding second-order derivatives. Moreover, Eq. (4.150) can be integrated giving

$$B'(u) \equiv (A/r^2)' = 2(3m - u)/r^4, \tag{4.152}$$

where $B(u) \equiv A/r^2$ and m is an integration constant equal to the Schwarzschild mass if the metric (4.124) is asymptotically flat as $u \to \infty$ ($r \approx u$, $A = 1 - 2m/u + o(1/u)$). If there is a flat asymptotic as $u \to -\infty$, the Schwarzschild mass there is equal to $-m$ ($r \approx |u|$, $A = 1 + 2m/|u| + o(1/u)$.

One more observation is that if the system contains a horizon and $r(u) \sim |u|$ at large $|u|$ in the T-region, then $B \equiv A/r^2$ tends to a finite limit, which means that there is a de Sitter asymptotic. Indeed, under these conditions the integral of (4.152) evidently converges at large $|u|$, so B tends to a constant. Furthermore, Eq. (4.150) can be rewritten as $r^4 B'' + 4r^3 r' B' = -2$, hence $B'' < 0$ where $B' = 0$, so $B(u)$ cannot have a minimum (and it is this circumstance that restricts the possible kinds of global causal structure of any scalar-vacuum solutions [53]). This means that $B(u)$ can only tend to a negative constant in a T-region.

Thus, in the Einstein-scalar system (4.147)–(4.151), *any* solution with a horizon and $r \sim |u|$ as $u \to \pm\infty$ is asymptotically de Sitter in the T-region. It describes a black universe if it is, in addition, asymptotically flat in the R-region.

As before, to find examples of solutions possessing particular properties, we employ the inverse problem method, choosing some of the functions $r(u)$, $A(u)$ or $\phi(u)$ and then reconstructing the form of $V(\phi)$ and/or $h(\phi)$. We will first choose a function $r(u)$ that can provide a black universe solution. Then $A(u)$ is found from (4.152) and $V(u)$ from (4.148). The function $\phi(u)$ is found from (4.149) provided $h(\phi)$ is known; however, using the scalar field parametrization freedom, we can, vice versa, choose a monotonic function $\phi(u)$ (which will yield an unambiguous function $V(\phi)$) and then find $h(u)$ from Eq. (4.149).

A simple example of the function $r(u)$ satisfying the requirements 1–3 is [88] (see Fig. 4.9):

$$r(u) = a\frac{(x^2+n)}{\sqrt{x^2+n^2}}, \quad n = \text{const} > 2. \tag{4.153}$$

where $x = u/a$, and $a > 0$ is an arbitrary constant (the minimum radius).

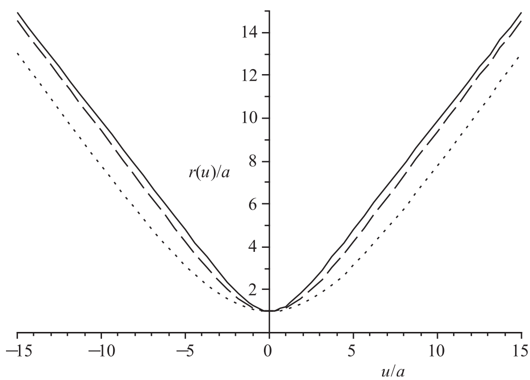


Figure 4.9 Plots of $r(u)/a$ given by Eq. (4.153) with $n = 3, 5, 10$ (solid, dashed, and dotted lines, respectively).

Integrating Eq. (4.152) under the assumption $m > 0$, we find (see Fig. 4.10)

$$B(u) = \frac{3x^4 + 3x^2 n(n+1) + n^2(n^2+n+1)}{3a^2(x^2+n)^3} + \frac{mx}{8a^2 n(x^2+n)^3}\Big[3x^4(5n^2+2n+1) + 8nx^2(5n^2+2n-1) + 3n^2(11n^2-2n-1)\Big] - \frac{3m}{8a^2 n^{3/2}}(5n^2+2n+1)\cot^{-1}\left(\frac{x}{\sqrt{n}}\right). \qquad (4.154)$$

The emerging integration constant is excluded by the requirement $B \to 0$ as $u \to \infty$, providing asymptotic flatness. Examples of the behavior of $B(u)$ for $m = 0.2a$ and some values of the parameter n are presented in Fig. 4.10.

Substituting the expressions (4.153) and (4.154) into (4.148), taking into account that $A(u) = B/r^2$, we obtain the potential V as a function of u or $x = u/a$. This expression is bulky and will not be presented here.

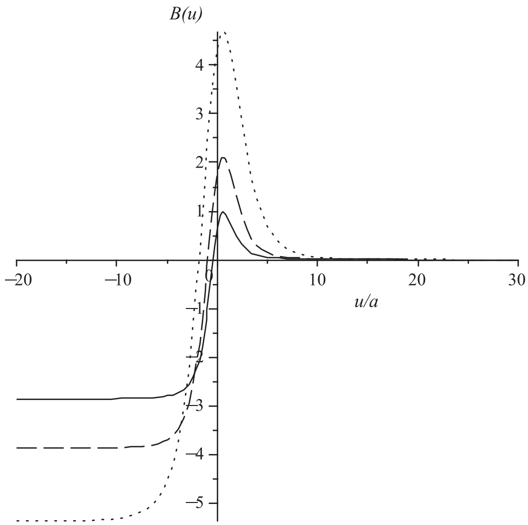


Figure 4.10 Plots of $B(u)$ given by Eq. (4.154) for $m = 0.2a$, $n = 3,\ 5, 10$ (solid, dashed, and dotted lines, respectively).

To construct V as an unambiguous function of ϕ and to find $h(\phi)$, it makes sense to choose a monotonic function $\phi(u)$. It is convenient to assume

$$\phi(u) = \frac{2\phi_0}{\pi} \arctan \frac{x}{n}, \quad \phi_0 = \frac{\pi a}{2} \sqrt{\frac{2(n-2)}{n}}, \tag{4.155}$$

and ϕ has a finite range: $\phi \in (-\phi_0, \phi_0)$, which is common to kink configurations. Thus we have $x = u/a = n \tan(\pi\phi/2\phi_0)$, whose substitution into the expression for $V(u)$ gives $V(\phi)$ defined in this finite range. The function $V(\phi)$ can be extended to the whole real axis, $\phi \in \mathbb{R}$, by supposing $V(\phi) \equiv 0$ at $\phi \geq \phi_0$ and $V(\phi) = V(-\phi_0) > 0$ at $\phi < -\phi_0$. Plots of $V(\phi)$ are presented in Fig. 4.11 for the same values of the parameters as in Fig. 4.10.

The expression for $h(\phi)$ is found from (4.149) as follows:

$$h(\phi) = \frac{(n-2)x^2 + n^2(1-2n)}{a^2(n-2)(x^2+n)}, \tag{4.156}$$

where $x = n \tan(\pi\phi/2\phi_0)$. The function $h(\phi)$ given by Eq. (4.156) is also defined in the interval $(-\phi_0, \phi_0)$ and can be extended to $\mathbb{R}$ by supposing $h(\phi) \equiv 1$ at $|\phi| \geq \phi_0$. The extended kinetic coupling function $h(\phi)$ is

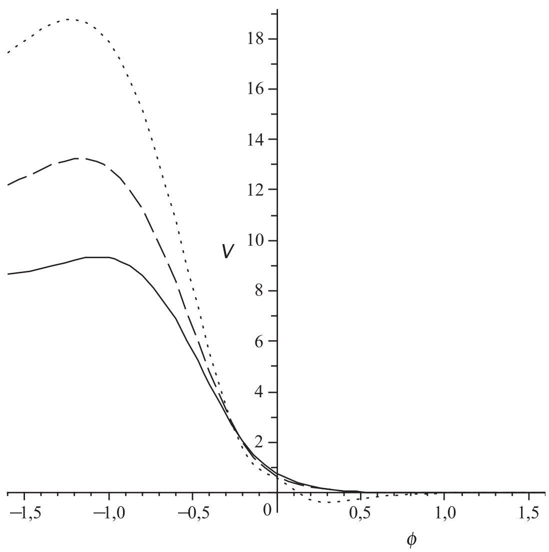


Figure 4.11 Plots of $V(u)$ for $m = 0.2a$ and $n = 3, 5, 10$ (solid, dashed, and dotted lines, respectively).

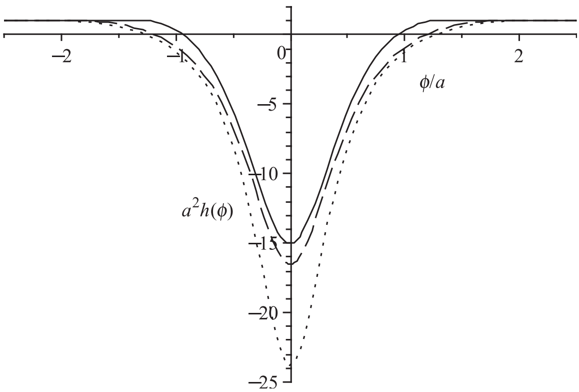


Figure 4.12 Plots of $h(\phi)$ given by Eq. 4.156 with $n = 3, 5, 10$ (solid, dashed, and dotted lines, respectively).

plotted in Fig. 4.12. Evidently, the null energy condition is violated only where $h(\phi) < 0$.

Thus we have shown [64, 88] that a minimally coupled scalar field may change its nature from canonical to ghost in a smooth way without creating any space-time singularities. This feature, in particular, allows for construction of wormhole models (trapped-ghost wormholes [88]) where the ghost is present in some restricted region around the throat (of arbitrary size) whereas in the weak-field region far from it the scalar has usual canonical properties. The same model allows for construction of black universes.

It has also been found that, in the Einstein-scalar field system under study, a static, spherically symmetric configuration is inevitably a black universe if it is asymptotically flat, has a horizon, and the function $r(u)$ grows linearly as $u \to \pm\infty$. However, all this can only happen if the scalar is of ghost nature at least in some part of space.

This page intentionally left blank

Chapter 5

Wormholes

5.1 The notion of a wormhole

In this chapter we will discuss one more type of configurations involving a significant space-time curvature, the wormholes. Above all, let us fix the terminology.

The term "wormhole" was introduced by J.A. Wheeler [422] in 1955. Wormholes are generally understood as comparatively narrow "bridges", or "tunnels" between comparatively large or infinitely extended space-time regions. Needless to say such an opportunity is widely used by authors of science fiction whose heroes easily and without much emotion leave behind them interstellar or thousands of years.

The possible appearance of a wormhole geometry is illustrated by Fig. 5.1, which presents two-dimensional images instead of three-dimensional ones. Each of the surfaces depicted contains a part having the shape of a thin tube that joins either flat "universes" (a) or spherical ones (b) (a dumbbell-like configuration), or a flat one with a spherical one (c) (a hanging-drop configuration), or distant regions of the same "universe" (d and e).

We see that the notion of a wormhole has in essence an intuitive nature since it rests on such words as "comparatively narrow" etc., but it contains an element that admits a rigorous definition. It is a throat, defined as a two-dimensional surface of minimal area, which does not admit a further contraction. In Fig. 5.1, the throats are one-dimensional circles of

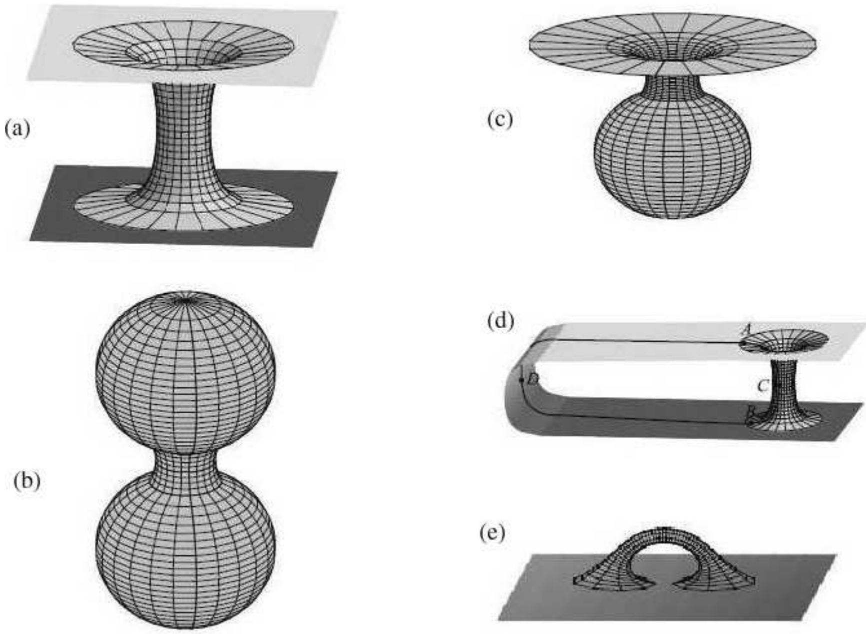


Figure 5.1 Diverse wormhole configurations, two-dimensional analogues.

minimum radius, which conditionally depict two-dimensional spheres of minimum area. It is obvious, however, that in general a wormhole need not be spherically symmetric, and its throat is not necessarily a sphere.

One should also add that all these shapes correspond to spatial sections of the four-dimensional space-time and, as usual, these spatial sections can not only evolve with time but can in general also depend on the choice of reference frame (recall the discussion of relativity of the spatial section topology in Sec. 2.2).

In the case of static, spherically symmetric space-times described by the general metric (3.1), it is not difficult to formulate the common requirements to be imposed on the metric coefficients which, if satisfied, enable one to say that this metric describes a wormhole geometry.

The condition that there are two large regions separated by a throat means that the function $r \equiv e^{\beta}(u)$, which is equal to the curvature radius of the coordinate sphere $u = \text{const}$, $t = \text{const}$ at each value of u, should have a regular minimum $r = r_{\min} > 0$ at some value of u and assume values $r \gg r_{\min}$ far from this minimum.

Another condition is that the wormhole must be freely traversable in both directions. So the existence of a proper three-geometry is insufficient: one must also require the absence of horizons. It is a fundamental distinction from black holes. For the metric (3.1), this requirement reads $g_{tt} \equiv \mathrm{e}^{2\gamma} > 0$ in the whole range of the radial coordinate.

The simplest example of a wormhole metric is

$$ds^2 = dt^2 - du^2 - (u^2 + a^2)d\Omega^2, \quad a > 0, \tag{5.1}$$

it differs from the Minkowski flat-space metric only by the emergence of the term a^2. But this drastically changes the geometry: at $a = 0$ we have the Minkowski space, and the coordinate u changes from zero to infinity, $u=0$ is a regular centre, and an extension to negative values of u is meaningless. However, if $a > 0$, the value $u = 0$ is a regular minimum of the function $r(u) = \sqrt{u^2 + a^2}$, i.e., a throat, nothing prevents consideration of $u < 0$, and $u \to -\infty$ is again a spatial infinity of the same kind as that at $u \to +\infty$. The 3-geometry corresponds to Fig. 5.1 and is time-independent.

To illustrate the importance of the second condition, let us make a small historical digression.

Right after the advent of general relativity, the researchers began to reflect on what could be expected if space-time is not simply curved but is strongly curved. As early as in 1916, the birth year of GR, Schwarzschild found his famous solution, Eq. (3.20), which describes, in modern terms, the simplest black hole. Less well known is the work of Ludwig Flamm [163] also dated 1916, where he noticed that the Schwarzschild solution describes something like a bridge or shortcut between two worlds or two parts of the same world, i.e., just what modern authors call a wormhole.

Let us explain this observation by passing on from the curvature coordinates in the metric (3.20) to the isotropic coordinates (3.2) by the transformation

$$r = u\left(1 + \frac{m}{2u}\right)^2. \tag{5.2}$$

As a result, the Schwarzschild metric acquires the form

$$ds^2 = \frac{(2 - m/u)^2}{(2 + m/u)^2} dt^2 - \left(1 + \frac{m}{2u}\right)^4 (du^2 + u^2 d\Omega^2), \tag{5.3}$$

where the coordinate $u \approx r$ as $r \to \infty$ and has the value $u = m/2$ at the horizon $r = 2m$. Here, it is convenient to introduce the dimensionless variable $y = 2u/m$. Then

$$ds^2 = \frac{(y-1)^2}{(y+1)^2}dt^2 - \frac{m^2}{4}\frac{(y+1)^4}{y^4}(dy^2 + y^2 d\Omega^2), \tag{5.4}$$

and the horizon corresponds to $y = 1$. The R-region $r > 2m$ is now described as the range $y > 1$. It is, however, easy to verify that the metric (5.4) does not change at an inversion, $y \mapsto 1/y$. It means that the range $0 < y < 1$ describes one more R-region, with quite the same geometry as the region $y > 1$, and the value $y = 0$ corresponds to another spatial infinity quite similar to the one at ($y = \infty$); moreover, the value $y = 1$, where the coordinate spheres have a minimum area equal to $16\pi m^2$, describes a throat in the spatial section of the 4-manifold under study. If the geometry were restricted to the spatial section only, we would have all reasons to say that there is a wormhole. However, everything changes due to $g_{tt} = 0$ at $y = 1$: as we know from the previous chapters, it is a horizon, whose existence means that the 4-geometry is inextendible in the static reference frame. For a more complete description one can introduce, e.g., the Kruskal coordinates, in which the complete geometry of the Schwarzschild space-time can be described. The resulting complete picture implies that the metric (5.4) describes only the two R-regions represented by the left and right quadrants in the Kruskal diagram. (Compare: the metric (3.20) at $r > 2m$ describes only one of the R-regions.) The sphere $y = 1$, which seems to be a boundary between the two R-regions when we look at Eq. (5.4), turns out to be a pair of different surfaces in the full description, and there is a whole T-region between them. The paradox is explained, in particular, by the fact that the horizon $r = 2m$ (or $y = 1$) *does not belong* to any of the R-regions, hence the unity of their description by the metric (5.4) is only seeming, and the coordinate regions $y > 1$ and $y < 1$, corresponding to different R-regions of the Schwarzschild space-time, should be considered separately.

Metrics like (5.4), containing horizons, are sometimes said to describe “nontraversable wormholes”. In our view, in such cases, to avoid confusion, one should speak of quite a different class of relativistic objects, namely, black holes. Even more than that, as we have seen while building the Carter-Penrose diagrams, *all* black holes with simple horizons (and any other horizons of odd order) contain quite the same pairs of R-regions as the Schwarzschild black hole, and all of them could be called “nontraversable wormholes” on the same grounds (more precisely, on the same level of uselessness).

It should be noted, though, that some families of solutions to the gravitational field equations can contain metrics describing objects of both classes at different values of the parameters, see, e.g., [60].

5.2 A wormhole as a time machine

The works on wormhole physics are certainly only theoretical; others are impossible right now. They can be divided into two trends, supplementing each other. In one of them, the researchers assume the existence of wormholes from the very beginning and try to infer what can come out of that. An important part of this trend is what can be called wormhole astrophysics: assuming that there are wormholes of astrophysically relevant size, it is reasonable to seek their observational signatures. In the other trend, the researchers (quite frequently, the same authors) try to find out under which conditions, natural or artificial, wormholes can emerge, and what will be their properties and peculiarities. Among such works are certainly numerous attempts to obtain particular wormhole solutions to the gravitational field equations.

Let us here dwell upon some questions belonging to the first trend, leaving aside for a while the question of "what the wormholes can be made of". We will discuss the (probably) most interesting version of a wormhole assuming that it connects different regions of the same space-time. To begin with, let us suppose that there is a stable wormhole somewhere in space, and that the space-time outside it is flat or almost flat, where, as usual, the laws of special relativity hold. Then, simply from the symmetries of Minkowski space it is clear that a wormhole can not only be an intergalactic tunnel but also an accelerator and even a time machine.

Indeed, the wormhole has two flat asymptotic regions. In each of them one can choose a reference frame (RF) (for simplicity, an inertial one) in which the wormhole mouth is at rest. But it follows from nowhere that it is the same RF for both mouths! And even if it is the same RF, and in the regions where the mouths are located, the clocks are synchronized in some reasonable way, it follows from nowhere that after passing through the wormhole the traveller will get into the same epoch by this common time, since the Minkowski metric is invariant under time translations. In other words, even if we know where a given wormhole (say, found in space by chance) leads, we in general cannot know "to when" it leads.

The invariance of the Minkowski metric under boosts, i.e., transformations from one inertial RF to another, leads to uncertainty in the velocity of the mouth of exit with respect to the RF of the mouth of entry. Thus, a traveller which has passed through the wormhole can find himself in motion with an arbitrary relativistic velocity with respect to the initial RF and (which is more important if the mouths are far from each other) with respect to the ambient bodies.

Thus the appearance, role and function of a wormhole depends on the positions and motion of its mouths, not simply in space but in space-time.

Another statement of the problem is also of interest. Suppose we are able to create a wormhole here and now, with quite macroscopic parameters, so that one can pass through a wormhole from one mouth to the other for a few seconds, and the mouths themselves are placed relatively close to each other somewhere in circumterrestrial space. Is it possible to convert such a wormhole into a time machine? It is shown by Morris, Thorne and Yurtsever [302] how to do that: the idea is to leave one of the mouths, to be denoted A, at rest, while the other, B (which should in principle behave as an ordinary massive body) should be accelerated to velocities comparable with the speed of light, then returned back and stopped near A. Then, due to the familiar special-relativistic effect, the Lorentzian time slowing-down on the moving body as compared with the one at rest, the time elapsed at B is smaller than at A. The particular difference depends on the velocity reached and the travel duration. It is actually just the well-known twin paradox: the twin returning from an interstellar trip at relativistic speed turns out to be younger than his brother sitting at home. If the difference between the mouths is, say, six months, then, being located near mouth A in January 2012, we would be able to see the green leaves of summer 2011 through the wormhole — and will really return there if we pass it through! If we again approach mouth A (as we agreed, it is not too far from B) and passing through the wormhole once more, we would get into last year's snow of January 2011. And so on, until we get tired... Naturally, under the assumption that the wormhole is stable or we ourselves are able to maintain it in operation.

Let us quote the beginning of the article [302], which looks like a slogan of wormhole physics: "Normally theoretical physicists ask: 'What are the laws of physics?' and/or 'What do these laws predict about our Universe?' In this letter we ask, instead, 'What constraints do the laws of physics place on the activities of an arbitrarily advanced civilization?' This will lead to some intriguing queries about the laws themselves."

Soon after the works of Thorne's group (1988) appeared the paper by Frolov and Novikov [167] who showed that even without special effort a wormhole in an external gravitational field, as time passes, inevitably turns into a time machine. Indeed, suppose that one of the mouths (A) of a stable wormhole is located near the surface of a neutron star and the other (B) far from it. Then the time rate is slower near A and faster near B, this difference is stored, and eventually there should appear closed timelike trajectories threading the wormhole. The time machine is ready.

The possible existence of wormholes would substantially affect black hole physics as well, see a discussion in the same paper [167]. Thus, nothing prevents us from assuming that mouth A of a wormhole is located outside a black hole while the other, B, is inside the horizon. Such a wormhole may be used to rescue an astronaut who fell into the black hole as well as for exploring the inner properties of the hole itself. Strictly speaking, the black hole horizon is then not an event horizon any more.

We have seen that if there is a stable wormhole at our disposal, we can rather easily obtain a time machine. The above reasoning makes it possible to conclude that the laws of physics (among which we include the notion of gravity as space-time curvature), in principle, do not forbid the existence of wormholes and consequently closed timelike and null trajectories that violate the causality principle. As a result, in papers on theoretical physics one can meet discussions of paradoxes whose place could be only expected in science fiction. What about meeting another (third, tenth) copy of oneself? What about the well-known "grandmother paradox" — the opportunity to penetrate into the past and kill one's grandmother in her childhood (or, in a milder version, simply to marry the young grandmother, then a girl, instead of the young grandfather)? Then the hero himself — where could he come from?

One of the well-known ways to surmount such and other paradoxes involves the hypothesis of parallel worlds, also called the Multiverse hypothesis, which is closely related to Everett's many-worlds interpretation of quantum theory [158]. Quantum processes are known to be of probabilistic nature, and any of them can have a few outcomes. By Everett, all possible outcomes are realized, but each of them in its own universe. Thus any quantum event (including, by the way, all particle interactions in any part of the Universe) increases the number of parallel worlds. Being so exotic and evidently cumbersome, such an interpretation is free of logical contradictions and rather naturally explains not only the quantum paradoxes but also the above-mentioned "grandmother paradox": getting to the past, the traveller actually creates a new universe, where a girl looking like his grandmother in youth can exist, but she is not at all the mother of his mother...

Another possible solution (more economical but not so easy to understand) was suggested by Novikov as the "self-consistency principle" [314], according to which "something can be realized locally only if it is globally self-consistent". It means [192] that events in a time loop affect each other along a closed cycle, and not only the past determines the future, but the future takes part in forming the past, which, whatever strange it can seem, does not lead to a violation of the causality principle.

There is, though, the radically conservative viewpoint that nature does not admit and cannot admit time loops — something like a chronological censorship. As any censorship, it cannot cause sympathy with the authors and many of their compatriots. But to put it seriously, this hypothesis, like any other, needs proof on the basis of the first principles. (As is easy to understand, any hypothesis claiming that something cannot exist can be only rejected by experiment but not proved.)

5.3 Wormholes as solutions to gravitational field equations

The works belonging to the second trend are more numerous and are also of much interest: among them there are searches for specific wormhole models, discussions of their properties and peculiarities that determine whether or not they might be realized, what could be done with them and how to use them.

Above all, if wormholes or, wider, space-times containing time loops are not forbidden completely, they should be quite rare in Nature, otherwise they would be observed long ago. Thus if a theory pretends to describe Nature, it should contain a mechanism forbidding or at least strongly preventing the formation of macroscopic wormholes.

No doubt, GR pretends to describe reality. However, there are many solutions describing wormholes and other spaces with closed timelike loops. But all of them are as a rule considered to be unrealistic or, in a sense, not dangerous. Thus, a very interesting solution to the Einstein equations has been found by Gödel [174]: it is a homogeneous stationary universe, rotating as a whole. It contains closed timelike trajectories, but a calculation shows that the smallest total duration of such a loop is much larger that the age of our Universe.

Important constraints follow from the very structure of the Einstein equations. The idea is that if we substitute a metric, possessing some properties of interest, to the left-hand side of the equations, it becomes possible to find explicit constraints on the possible form of the right-hand side (that is, the SET) compatible with these properties. Let us look how it is done using, as a simple but important example, static, spherically symmetric space-times.

5.3.1 Spherically symmetric wormholes. General properties

Let us, as before, begin with the general form (3.1) of a static, spherically symmetric metric

$$ds^2 = \mathrm{e}^{2\gamma} dt^2 - \mathrm{e}^{2\alpha} du^2 - \mathrm{e}^{2\beta} d\Omega^2, \tag{5.5}$$

and consider some general properties of static, spherically symmetric wormholes, assuming the SET in the general form compatible with this kind of symmetry, viz.,

$$T_\mu^\nu = \mathrm{diag}(\rho,\ -p_r,\ -p_\perp,\ -p_\perp), \tag{5.6}$$

where $\rho,\ p_r,\ p_\perp$ are the density, radial pressure and transversal pressure respectively, which in general can be arbitrary functions of the radial coordinate u.

We will make use of two different radial coordinates, namely, the quasiglobal and curvature coordinates.

With the quasiglobal coordinate[17] u $(\alpha + \gamma = 0)$, the metric has the form (3.30),

$$ds^2 = A(u)dt^2 - \frac{du^2}{A(u)} - r^2(u)d\Omega^2, \tag{5.7}$$

and the conditions on a wormhole throat $u = u_{\rm th}$, expressing the absence of a horizon and a finite regular minimum of r, have the form $A > 0$,

[17] In this section we denote this coordinate by the letter u instead of ρ since the letter ρ denotes density.

$r > 0$, $r' = 0$, $r'' > 0$. Then two of the Einstein equations (see (3.5), (3.6))

$$R^0_0 - R^1_1 = 2A\frac{r''}{r} = -\varkappa(\rho + p_r), \tag{5.8}$$

$$G^1_1 = \frac{1}{r^2}[-1 + A'rr' + Ar'^2] = \varkappa p_r, \tag{5.9}$$

lead to the following conditions on the throat:

$$\rho + p_r < 0, \quad p_r < 0. \tag{5.10}$$

Notice that there is no restriction on the sign of the density. The conditions (5.10) are of local nature and do not depend on any assumptions on the space-time or matter properties far from the throat. It is clear that both inequalities are quite unusual for macroscopic matter, though for field matter the second condition (5.10) can be easily satisfied, e.g., for a radial electric field we have, according to (3.14), $p_r = -\rho$.

It is easy to verify that the first inequality (5.10) violates one of the most important conditions, satisfied by the majority of known kinds of matter. It is the Null Energy Condition (NEC). In the general case, this condition is expressed by the inequality

$$T_{\mu\nu}\xi^\mu\xi^\nu \geq 0, \quad \forall\, \xi^\mu : \quad \xi^\nu\xi_\mu = 0. \tag{5.11}$$

In the metric (5.5), for the null vector $\xi^\mu = (\,\mathrm{e}^{-\gamma},\ \mathrm{e}^{-\alpha},\ 0,\ 0)$, this condition takes the form

$$T^0_0 - T^1_1 \geq 0. \tag{5.12}$$

For the tensor (5.6) it reads $\rho + p_r \geq 0$, contrary to (5.10). Thus the wormhole existence requires NEC violation.

Matter that violates the NEC is commonly called exotic. The conclusion that it is this kind of matter that is necessary for wormhole existence is not restricted to cases that are static or spherically symmetric and has quite a general nature: it holds true for throats of any spatial configuration, both stationary and time-dependent [204, 205] (though for the latter even the definition of a throat is rather complicated [205]).

But let us return to static spherical symmetry and now write the metric in terms of the curvature coordinates:

$$ds^2 = \mathrm{e}^{2\gamma(r)}dt^2 - \mathrm{e}^{2\alpha(r)}dr^2 - r^2 d\Omega^2. \tag{5.13}$$

Identifying the line elements (5.7) and (5.13), it is not hard to verify that on a throat, if any, the following conditions hold: $|\gamma| < \infty$, $\mathrm{e}^{-2\alpha} \equiv B(r) = 0$,

$dB/dr \neq 0$. One of the Einstein equations (3.5) has the form

$$-G_0^0 = \frac{1}{r^2}(1 - \mathrm{e}^{-2\alpha}) + \frac{2\alpha' \, \mathrm{e}^{-2\alpha}}{r} = \varkappa\rho \qquad (5.14)$$

and can be integrated: putting $\mathrm{e}^{-2\alpha} = 1 - 2m(r)/r$, we bring it to the form $2m'(r) = \varkappa\rho r^2$ and consequently,

$$m(r) = 4\pi G \int_{r_0}^{r} \rho r^2 dr, \qquad (5.15)$$

where r_0 is an integration constant. The function $m(r)$ is called the mass function; in the case of an asymptotically flat configuration with a regular centre it really gives the Schwarzschild mass if one integrates from zero to large radii where the metric is approximately Schwarzschild. However, the integral (5.15) is quite general for the metric (5.13).

In the case of an asymptotically flat wormhole, integrating in (5.15) from the throat $r = r_{\mathrm{th}}$ to large values of the radius, we obtain

$$r_{\mathrm{th}} = 2m - \varkappa \int_{r_{\mathrm{th}}}^{\infty} \rho r^2 dr, \qquad (5.16)$$

since on the throat $B(r) \equiv 1 - 2m(r)/r = 0$. Recalling that $2m(\infty) = 2m$ is the Schwarzschild radius corresponding to the mass m, we come to the conclusion that *the radius of a wormhole throat is larger than the Schwarzschild radius corresponding to the wormhole mass at its flat asymptotic only if the density of matter supporting the wormhole is negative; the throat radius is smaller than the Schwarzschild one if the density is positive.*

It is really an important constraint: indeed, if we wish to have a wormhole with a throat of a few meters or kilometers rather than something of planetary or stellar size, we must take into account that meters and kilometers are just the scale of planetary and stellar Schwarzschild radii (for the Earth it is of the order of 1 cm, for the Sun about 3 km). To avoid huge gravity near the wormhole mouth and/or throat corresponding to such a mass, we must invoke not simply exotic matter, but matter with negative density.

5.3.1.1 *Wormholes with scalar fields*

We will now consider a few examples of exact wormhole solutions to the gravitational field equations.

Wormholes in the anti-Fisher solution

The simplest example is given by the anti-Fisher solution described by Eq. (4.30) with $\varepsilon = -1$: in the case $k < 0$ it describes a wormhole with two spatial asymptotics at $u = 0$ and $u = \pi/|k|$. The metric has the form

$$\begin{aligned} ds^2 &= \mathrm{e}^{-2hu} dt^2 - \frac{k^2\,\mathrm{e}^{2hu}}{\sin^2(ku)}\left[\frac{k^2\,du^2}{\sin^2(ku)} + d\Omega^2\right] \\ &= \mathrm{e}^{-2hu} dt^2 - \mathrm{e}^{2hu}[d\rho^2 + (k^2+\rho^2)d\Omega^2], \end{aligned} \tag{5.17}$$

where the harmonic coordinate u is expressed in terms of the quasiglobal coordinate ρ, specified on the whole axis $\mathbb{R}$, by $|k|u = \arctan(\rho/|k|)$.

At $m > 0$, the wormhole attracts the ambient test matter on the first asymptotic ($\rho \to \infty$) and repels it on the other ($\rho \to -\infty$), and vice versa in the case $m < 0$. At $m = 0$ we obtain the simplest solution called the Ellis wormhole (although H. Ellis's paper [155] discussed anti-Fisher wormholes of any mass).

The wormhole throat is located at $\rho = h$ and has the size

$$r_{\rm th} = (h^2 + k^2)^{1/2} \exp\left(\frac{h}{k}\cot^{-1}\frac{h}{k}\right). \tag{5.18}$$

It has been shown in the recent paper [175] that the anti-Fisher wormholes are unstable under small radial perturbations, and the characteristic time of perturbation growth is of the order $r_{\rm th}/c$, i.e., the time needed for a light beam to cover a distance of the order of the throat size.

Analogues of the metric (5.17) with a nonzero electric charge are also known [50]; they can be obtained from (4.24), (4.25) at $\varepsilon = -1$: these are counterparts of Penney's solution [324] with a phantom scalar field. The dependence of their stability or instability on the charge value has not been studied yet. The stability of static, spherically symmetric configurations will be discussed in the next chapter.

Quite naturally, wormholes can also be supported by minimally coupled phantom scalar fields with nonzero potentials, see, e.g., [66]; a more unusual observation is that wormhole solutions can be obtained with the "trapped ghosts" [88], scalar fields which exhibit phantom properties only in a string-field restricted region of space whereas in the rest of space all standard energy conditions are observed. Such wormholes can only exist with nonzero scalar field potentials. An example of such a solution can be found among the solutions presented in Sec. 4.5.4: it corresponds to $m = 0$.

Wormholes with nonminimally coupled scalar fields

In the case of a conformal scalar field with the Lagrangian (4.1), (4.4), there exists a family of solutions to the field equations having the form (4.32)–(4.33), with the metric [50]

$$ds^2 = (y + y_0)^2 \left[\frac{dt^2}{(y + y_1)^2} - h^2 \frac{(y + y_1)^2}{y^4} (dy^2 + y^2 d\Omega^2) \right]; \quad (5.19)$$

$$y_0 = \tanh hu_0, \quad y_1 = \coth hu_1, \quad (5.20)$$

and the scalar field

$$\phi = \sqrt{6} \frac{1 + yy_0}{y + y_0}. \quad (5.21)$$

In the electrically neutral case $Q = 0$ we have $y_1 = 1$; spatial infinity corresponds to $y = \infty$, where the scalar field tends to $\phi = \phi_\infty = \sqrt{6} y_0$.

The metric (5.19) describes a wormhole in the case $y_0 > 0$; the coordinate y then has the range $y > 0$. It is easy to verify that as $y \to 0$, as well as for $y \to \infty$, the metric becomes flat. Thus $y = 0$ is the second spatial infinity, where $e^{2\gamma} \to (y_0/y_1)^2$ and $\phi \to (1/y_0)\sqrt{6}$. Thus we have a wormhole with a throat at $y = \sqrt{y_0 y_1}$.

The metric (4.26) (in other notations) and its generalizations to other nonminimally coupled scalar fields (the Lagrangian (4.1), $f(\phi) = 1 - \xi\phi^2$, $h(\phi) = 1$, where $\xi > 0$ is an arbitrary constant) were discussed by Barcelo and Visser [21, 22] as examples of wormholes existing even with a scalar field with the normal sign of kinetic energy.

However, all wormholes supported by nonminimally coupled scalar fields and, more generally, wormholes existing due to conformal continuations in scalar-tensor theories [54] have a common peculiarity: at the transition sphere, at which the nonminimal coupling function $f(\phi)$ in the Lagrangian (4.1) passes through zero, the effective gravitational constant also changes its sign. As a result, the second spatial infinity of such wormholes is situated in a region with a negative effective gravitational constant (this phenomenon is sometimes characterized by the words "the graviton is there a ghost"). The plausibility of such configurations is thus questionable; also, evidently, they cannot connect distant regions of the same universe, and their remote mouth should be situated in a world where the physical laws are very unusual from our point of view. Besides, it has been established [70, 71] that electrically neutral (or weakly charged) wormholes of this kind are unstable under spherically symmetric perturbations (see also Chapter 6), and there are reasons to believe that such an instability has a common nature since it is related to the field behaviour near the transition surface with $f = 0$.

(For more details on conformal continuations in static, spherically symmetric configurations in scalar-tensor and $f(R)$ gravity, the necessary and sufficient conditions for their existence and the possible full geometries of the continued space-times see [54–56].)

More generally, in the framework of STT and $f(R)$ theories of gravity (for instance, in the Brans-Dicke theory), wormhole solutions evidently do exist and can be obtained from those of GR with a phantom scalar field with the aid of the conformal mapping (4.3) provided the conformal factor $f(\phi)$ is regular everywhere. Thus, all wormhole solutions with a massless Brans-Dicke scalar are transforms of the anti-Fisher wormholes discussed above, except for the special case $\omega = 0$ [51] where a special wormhole solution exists due to a conformal continuation (see Sec. 5.3.3).

5.3.2 Wormhole construction by solving the trace of the Einstein equations

In this section we demonstrate the significant level of arbitrariness in wormhole solutions of the Einstein equations if we do not specify the form of exotic matter from the beginning. We will show that even if, for simplicity, we restrict ourselves to matter with zero trace of the SET, the wormhole solutions form a set parametrized by one arbitrary function plus an integration constant [87].

We will start with the general static, spherically symmetric metric in 4 dimensions written in the curvature coordinates

$$ds^2 = \mathrm{e}^{2\gamma(r)}dt^2 - \mathrm{e}^{2\alpha(r)}dr^2 - r^2 d\Omega^2, \qquad (5.22)$$

where $d\Omega^2 = d\theta^2 + \sin^2\theta\, d\phi^2$ is the linear element on a unit sphere.

If the SET of matter has a zero trace, the equation $R = 0$ is valid, and it can be written as a linear first-order equation with respect to $F(r) := r\,\mathrm{e}^{-2\alpha}$

$$F_r(2 + r\gamma_r) + F(2r\gamma_{rr} + 2r\gamma_r^2 + 3\gamma_r) = 2, \qquad (5.23)$$

where the subscript r means d/dr.

Let us note that Eq. (5.23) is of particular interest because the trace equation is the only uniquely determined equation describing the 4D gravitational field in an RS type brane world since the set of equations for gravity on the brane [371] contains a contribution $E_{\mu\nu}$ depending on the

5D Weyl tensor, having zero trace. (The basic concepts of the brane-world scenario are briefly discussed in Sec. 4.6.3.) In vacuum, where matter on the brane is absent and the 4D cosmological constant is zero (a natural assumption for scales much smaller than the size of the Universe), these equations reduce to

$$G_{\mu\nu} = -E_{\mu\nu}, \tag{5.24}$$

where $G_{\mu\nu}$ is the 4-dimensional Einstein tensor corresponding to the brane metric $g_{\mu\nu}$ while $E_{\mu\nu}$ is the projection of the 5-dimensional Weyl tensor onto the brane. The traceless tensor $E_{\mu\nu}$ connects gravity on the brane with the bulk geometry (and is sometimes called the tidal SET), so that the set of equations (5.24) is not closed. Due to its geometric origin, $E_{\mu\nu}$ does not necessarily satisfy the energy conditions applicable to ordinary matter. Thus, examples are known [408] when negative energies on the brane are induced by gravitational waves or black strings in the bulk. Therefore, if the brane world concept is taken seriously, $E_{\mu\nu}$ can be the most natural "matter" supporting wormholes.

Thus the wormhole metrics obtained with the algorithms to be described here may be considered as vacuum wormhole metrics in a brane world [87].

The general solution of Eq. (5.23) is

$$F(r) = \frac{2\,\mathrm{e}^{-2\gamma+3\Gamma}}{(2+r\gamma_r)^2}\int (2+r\gamma_r)\,\mathrm{e}^{2\gamma-3\Gamma}\,dr, \tag{5.25}$$

where

$$\Gamma(r) = \int \frac{\gamma_r dr}{2+r\gamma_r}. \tag{5.26}$$

Thus, choosing the form of $\gamma(r)$ arbitrarily, we obtain $F(r)$ from (5.25), and, after fixing the integration constant, the metric is known completely at least in the region where e^{γ} and e^{α} are smooth and nonzero.

From the solution (5.25), wormhole metrics (as well as black hole metrics [60]) are built algorithmically by specifying the generating function $\gamma(r)$ with desirable properties.

Let us note for reference purposes that in many articles on wormholes, beginning with [302], the function $\mathrm{e}^{2\alpha}$ in the curvature coordinates is written in the form $[1 - b(r)/r]^{-1}$, where $b(r)$ is called "the shape function". This name is, in our view, not quite appropriate since $b(r)$ bears information on the wormhole profile in a very indirect form; one could more reasonably call $r(l)$ in the metric (5.27) a shape function.

In fact (see (5.15)) $b(r) = 2m(r)$, where $m(r)$ is the mass function. The function $F(r)$ introduced here is $F(r) = r - b(r) = r - 2m(r)$.

Let us now make clear how to choose the function $\gamma(r)$ (the so-called redshift function) and the integration constant in Eq. (5.25) in order to obtain a wormhole solution.

The coordinate r, which proves to be convenient for solving Eq. (5.23), is not an admissible coordinate in the whole space for wormhole solutions since in this case r has at least one minimum, and the solution in terms of r therefore splits into at least two branches. As an admissible coordinate one can take, e.g., the Gaussian coordinate l (proper length along the radial direction) connected with r by the relation $l = \int e^{\alpha} dr$, and the metric is rewritten as

$$ds^2 = e^{2\gamma(l)} dt^2 - dl^2 - r^2(l) d\Omega^2. \tag{5.27}$$

We seek static, traversable, twice asymptotically flat wormhole solutions. So we require:

1. There should be two flat asymptotics: $l \in \mathbb{R}$; $r \approx |l| \to \infty$ and $\gamma = \text{const} + O(r^{-1})$ as $l \to \pm\infty$;
2. Both functions $r(l) > 0$ and $\gamma(l)$ should be smooth (at least C^2) in the whole range $l \in \mathbb{R}$.

This guarantees the absence of curvature singularities and horizons (the latter correspond to $\gamma \to -\infty$ which is ruled out). This also means that $r(l)$ should have at least one regular minimum, $r_{\min} > 0$ (throat), at some value of l. Moreover, returning to functions of r, we see that at a flat asymptotic $e^{\alpha} \to 1$ and $F(r) \approx r$.

Suppose, without loss of generality, that a minimum of $r(l)$, that is, a wormhole throat, is located at $l = 0$. Then $r(0) = r_0 > 0$, $r_l(0) = 0$ and (generically) $r_{ll}(0) > 0$, where the subscript l denotes d/dl. Near $l = 0$ one has $r - r_0 \sim l^2$, hence the metric function $e^{2\alpha(r)}$ behaves as $(r - r_0)^{-1}$, and $F(r) = r\, e^{-2\alpha} \sim r - r_0$. In other words, a simple zero of $F(r)$ is ***an indicator of a wormhole throat*** provided $\gamma(r)$ is smooth and finite at the same r.

On the other hand, the derivative $\gamma_l(0)$ may be zero (which is always the case if the wormhole is symmetric with respect to the throat) or nonzero. If $\gamma_l(0) = 0$, we shall have $\gamma_r(r_0) < \infty$. If, on the contrary, $\gamma_l(0) \neq 0$, then near r_0 we have $\gamma_r \sim 1/|l| \sim \sqrt{r - r_0}$, so that

$$\gamma(r) \approx \gamma(r_0) + k\sqrt{r - r_0}, \quad k > 0. \tag{5.28}$$

We cannot put $k < 0$ since then we would obtain the expression $2 + r\gamma_r$ ranging from 2 (at spatial infinity) to $-\infty$ at $r = r_0$, so that $2+r\gamma_r$ would vanish at some $r > r_0$ causing a singularity in (5.25).

We are now ready to single out a class of symmetric wormhole metrics (WH1) and a class of potentially asymmetric wormhole metrics (WH2) on the basis of the solution (5.25).

WH1. Specify the function $\gamma(r)$, smooth in the range $r_0 \le r < \infty$, $r_0 > 0$, in such a way that $\gamma(\infty) = 0$, $\gamma_r(r_0) < \infty$, and $2 + r\gamma_r > 0$ in the whole range. Fix the integration constant in (5.25) by performing integration from r_0 to r. Then these $\gamma(r)$ and $F(r)$ determine a wormhole which has a throat at $r = r_0$ and is symmetric with respect to the throat.

Indeed, by construction, $F(r) \sim r - r_0$ near r_0. Introducing the new coordinate x by the relation $r = r_0 + x^2$, we have $\mathrm{e}^{2\alpha} dr^2 \sim (r - r_0)^{-1} dr^2 = 4dx^2$, which leads to a perfectly regular metric whose coefficients are all even functions of $x \in \mathbb{R}$. Both $x \to +\infty$ and $x \to -\infty$ are flat asymptotics.

Each $\gamma(r)$ chosen as prescribed creates a family of symmetric wormholes with zero scalar curvature. The family is parametrized by the throat radius r_0, taking arbitrary values in the range where $\gamma(r)$ is regular and $2 + r\gamma_r > 0$.

Another procedure is applicable to functions $\gamma(r)$ behaving according to Eq. (5.28).

WH2a. Specify the function $\gamma(r)$, smooth in the range $r_0 \le r < \infty$, $r_0 > 0$, such that $\gamma(\infty) = 0$, $2+r\gamma_r > 0$ in the whole range, and Eq. (5.28) holds near r_0. Then, for proper values of the integration constant in (5.25), the sphere $r = r_0$ is a wormhole throat, and the solution is smoothly continued beyond it.

Indeed, the solution (5.25) may be rewritten as follows:

$$F(r) = \frac{\mathrm{e}^{-2\gamma+3\Gamma}}{\left(1 + \frac{1}{2} r\gamma_r\right)^2} \left[\int_{r_0}^{r} \left(1 + \frac{1}{2} r\gamma_r\right) \mathrm{e}^{2\gamma - 3\Gamma}\, dr + C \right]. \tag{5.29}$$

Suppose $C > 0$. Then $F(r)$ behaves near r_0 as $r - r_0 =: x^2$, while $\gamma = \gamma(r_0) + kx + O(x^2)$. The metric behaves smoothly at $r = r_0$ $(x = 0)$ in terms of the new coordinate x and can be continued through this sphere. One cannot, however, guarantee that this continuation will lead to another flat spatial infinity to yield an asymmetric wormhole, since the further behavior of $\gamma(x)$ and $F(x)$ may lead to a horizon or to a singularity.

If we choose $C \le 0$ in (5.29), we obtain two other situations:

WH2b. If $C < 0$, then $f(r_0) < 0$; recalling that $f \sim r$ at large r, we see that $F(r) = 0$ at some value $r = r_1 > r_0$, where γ_r is finite, and we return to the circumstances described as WH1, obtaining a symmetric wormhole with $r \ge r_1$, and the sphere $r = r_1$ is its throat.

WH2c. If $C = 0$, then near r_0 we obtain $F(r) \sim (r - r_0)^{3/2}$, and the metric is regularized at $r = r_0$ by another substitution: $r - r_0 = \xi^4$. As a result, Eq. (5.28) yields

$$\gamma = \gamma(r_0) + k\xi^2 + \text{further even powers of } \xi,$$

and we again obtain a symmetric wormhole, but now with a quartic behavior of r near its minimum as a function of the admissible coordinate $\xi \in \mathbb{R}$.

Now we will give some examples, presenting expressions for the metric functions γ and F, the effective "tidal" energy density ρ and the sum $\rho + p_{\rm rad}$, which characterizes violation of the Null Energy Condition (for static, spherically symmetric systems this condition reduces to $\rho + p_{\rm rad} \ge 0$).

We use the time scale of a distant observer at rest and so always assume that $e^\gamma \to 1$ as $r \to \infty$.

Example 1. The simplest example is obtained for $\gamma \equiv 0$. Choosing any $r_0 > 0$ and applying the W1 algorithm of Sec. 2, we simply obtain $F(r) = r - r_0$. This is a symmetric wormhole solution known as the spatial Schwarzschild geometry [123]:

$$\begin{aligned} ds^2 &= dt^2 - \left(1 - \frac{r}{r_0}\right)^{-1} dr^2 - r^2 d\Omega^2 \\ &= dt^2 - 4(r_0 + x^2)dx^2 - (r_0 + x^2)^2 d\Omega^2. \end{aligned} \tag{5.30}$$

The effective SET E_μ^ν has the form $T_\mu^\nu = \mathrm{diag}(0, -p_r, p_r/2, p_r/2)$ with the radial pressure

$$p_r = -r_0/r^3. \tag{5.31}$$

Example 2. Our next example uses the Schwarzschild form of γ:

$$e^{2\gamma} = 1 - \frac{2m}{r}, \quad m > 0. \tag{5.32}$$

Choosing any $r_0 > 2m$, we obtain according to the W1 prescription:

$$F(r) = \frac{(r-2m)(r-r_0)}{r-3m/2}, \tag{5.33}$$

$$\begin{aligned} ds^2 &= \left(1 - \frac{2m}{r}\right)dt^2 - \frac{1-3m/(2r)}{(1-2m/r)(1-r_0/r)}dr^2 - r^2 d\Omega^2 \\ &= \frac{x^2+r_0-2m}{x^2+r_0}dt^2 - \frac{4(x^2+r_0)(x^2+r_0-3m/2)}{x^2+r_0-2m}dx^2 \\ &\quad - (x^2+r_0)^2 d\Omega^2. \end{aligned} \tag{5.34}$$

This is evidently a symmetric wormhole geometry for any $r_0 > 2m \geq 0$, or for any $r_0 > 0$ in case $m < 0$. The Schwarzschild metric is restored from (5.34) in the special case $r_0 = 3m/2$.

The SET components of interest are

$$\rho = \frac{m(2r_0 - 3m)}{r^2(2r-3m)^2}; \quad \rho + p_{\rm rad} = -\frac{2(r-2m)(2r_0-3m)}{r^2(2r-3m)^2}. \tag{5.35}$$

The metric (5.34) was obtained by Casadio, Fabbri and Mazzacurati [104] in search for new brane-world black holes and by Germani and Maartens [171] as a possible external metric of a homogeneous star on the brane, but the existence of traversable wormhole solutions for $r_0 > 2m$ (in the present notations) was not mentioned. It was supposed in [104] that the post-Newtonian parameters of the metric must be close to their Einstein values for experimental reasons and therefore the authors restricted their study to configurations close to Schwarzschild. Then r_0 must be close to $3m/2$. In this case, as in the Schwarzschild metric, $r = 2m$ is an event horizon, but according to [104], the space-time structure depends on the sign of $\eta = r_0 - 3m/2$. If $\eta < 0$, the structure is that of a Schwarzschild black hole, but the curvature singularity is located at $r = 3m/2$ instead of $r = 0$. If $\eta > 0$, the solution describes a nonsingular black hole with a minimum of r at $r = r_0$ inside the horizon, which may be called a non-traversable wormhole [104]. This kind of configuration is mentioned in the classification of regular black holes in Sec. 4.6.1.

We would here remark that, in our view, such hypothetic objects as brane-world black holes or wormholes, not necessarily of astrophysical size, need not necessarily conform to the restrictions on the post-Newtonian parameters obtained from the Solar system and binary pulsar observations, and it therefore makes sense to discuss the full range of parameters which are present in the solutions.

Example 3. Consider the extreme Reissner-Nordström form of $\gamma(r)$:

$$e^{2\gamma} = \left(1 - \frac{2m}{r}\right)^2, \quad m > 0. \tag{5.36}$$

The W1 procedure now leads to

$$F(r) = \frac{(r - r_0)(r - r_1)}{r}, \quad r_1 \stackrel{\text{def}}{=} \frac{mr_0}{r_0 - m}; \tag{5.37}$$

$$\begin{aligned} ds^2 &= \left(1 - \frac{2m}{r}\right)^2 dt^2 - \left(1 - \frac{r_0}{r}\right)^{-1}\left(1 - \frac{r_1}{r}\right)^{-1} dr^2 - r^2 d\Omega^2 \\ &= \left(1 - \frac{2m}{x^2 + r_0}\right)^2 dt^2 - 4(r_0 + x^2)\left(1 - \frac{r_1}{r_0 + x^2}\right)^{-1} dx^2 \\ &\quad - (r_0 + x^2) d\Omega^2. \end{aligned} \tag{5.38}$$

where we assume $r_0 > 2m$, so that $r_1 < r_0$. It is a symmetric wormhole metric. The SET components of interest are

$$\rho = \frac{mr_0^2}{r^4(r_0 - m)}, \quad \rho + p_{\text{rad}} = -\frac{(r_0 - 2m)^2}{r^2(r - 2m)(r_0 - m))}. \tag{5.39}$$

In the solution (5.37), r_0 may be regarded as an integration constant, so it is of interest what happens if $r_0 \leq 2m$. Evidently, $r_0 = 2m$ leads to the extreme Reissner-Nordström black hole metric (which is well known to possess a zero Ricci scalar, as does the general Reissner-Nordström metric). In case $2m > r_0 > m$, we have $r_1 > 2m$, and we again obtain a symmetric wormhole, but now r ranges from r_1 to infinity and $r = r_1$ is the throat. Essentially, r_0 and r_1 exchange their roles as compared with the case $r_0 > 2m$. This property is expected due to symmetry between r_0 and r_1 in the metric (5.38).

The value $r_0 = m$ is meaningless. Lastly, $r_0 < m$ leads either to $r_1 < 0$ (for $r_0 \geq 0$) or to $0 < r_1 < 2m$ (for $r_0 < 0$). The solution exists in both cases for $r > 2m$ only, and $r = 2m$ turns out to be a naked singularity, as is confirmed by calculating the Kretschmann scalar.

We have seen that the equation $R = 0$ leads to a great number of wormhole solutions. Symmetric wormhole solutions of class WH1 can be obtained from any $\gamma(r)$ providing asymptotic flatness; asymmetric wormhole solutions belonging to class WH2a require somewhat more special conditions and are more difficult to obtain from the general solution (5.25); see an example in [87].

As is seen from Examples 1–3, wormholes are not always connected with negative (effective) energy densities ρ. They can appear with $\rho > 0$, but only with comparatively large negative pressures maintaining violation of the null energy condition; recall in this respect our general reasoning in Sec. 5.3.1. Example 3 shows that, for given $\gamma(r)$, sometimes even more wormhole solutions than expected can be obtained in search for class WH1 solutions.

All the above refers to metrics satisfying the condition $R=0$ in 4 dimensions, which admits an interpretation as the brane metric. We did not discuss the possible bulk properties of models in question; it can be noted that the existence of the corresponding solutions to the 5D equations of gravity (in our case, the 5D vacuum Einstein equations with a cosmological constant) is guaranteed by the Campbell-Magaard type embedding theorems [98, 286]. A recent discussion of these theorems, applied in particular to brane-world scenarios, and more references can be found in Ref. [366]. It has also been claimed that "any 4D space-time with $R = 0$ gives rise to a 3-brane world without surface stresses embedded in a 5-dimensional space-time" [409] since the embedding contains a very significant arbitrariness. Nevertheless, a complete model requires knowledge of the full 5D space-time. In other words, one should "evolve" the 4-metric into the bulk, using this 4-matric as initial data for the 5-dimensional equations. It is rather a difficult task, as was demonstrated in a study of particular black hole solutions in Refs. [104, 107]. There are, however, two favorable circumstances. One is the wealth of wormhole solutions: there is actually an arbitrary function $\gamma(r)$ leading to wormholes on the brane, which must in turn lead to a wide choice of suitable bulk functions. The other is the global regularity of wormhole space-times, and one can expect that the bulk incorporating them will also be regular. (It may be recalled that it was the singular nature of black hole solutions that caused some technical difficulties in Ref. [107].)

5.3.3 Alternative gravity and vacuum as wormhole supporters

The algorithms presented in the previous subsection demonstrate the wealth and diversity of wormhole metrics even in the narrow class of spherically symmetric space-times. Still this does not remove the question of possible real material sources of wormhole geometry satisfying, under the same symmetry, the conditions (5.10).

As we have seen, usual classical (nonphantom) matter, including fields, does not lead to wormholes since it respects the NEC (5.11). This result is true in the Einstein theory, but it extends to a large class of alternative theories of gravity where the gravitational equations are reduced to those of GR with an additional nonphantom scalar field. These are the scalar-tensor theories (STT) under the condition $\varepsilon = +1$ (see Sec. 4.1, Eqs. (4.1)–(4.3)), $F(R)$ theories, which have recently become very popular and where, instead of the GR Lagrangian ($\text{const} \cdot R$), the gravitational Lagrangian is a certain function $F(R)$, and the class of theories that unifies the STT and $F(R)$ theories, where the Lagrangian contains a function of two variables $F(\phi, R)$. The transformation from the initial equations involving the physical metric $g_{\mu\nu}$ (the so-called Jordan picture) to the equations of GR with minimally coupled scalar fields (the Einstein picture) have been obtained in a general form in the papers [412] for STT, [25, 26, 148, 228, 428] for $F(R)$ theories and [284] for $F(\phi, R)$ theories.

Indeed, let the metric $g_{\mu\nu}$ in the initial manifold $\mathbb{M}_\text{J}$ (the Jordan picture) and the metric $\overline{g}_{\mu\nu}$ in the manifold $\mathbb{M}_\text{E}$ (the Einstein picture) be related by the conformal mapping

$$g_{\mu\nu} = f(x)\overline{g}_{\mu\nu}, \tag{5.40}$$

where $f(x)$ is an everywhere regular function of the space-time coordinates, bounded both above and below by some positive constants. According to [204], in GR, for the existence of a static throat (defined as a minimal 2-surface in 3-space) it is necessary that the material source in the Einstein equations violate the NEC. Thus, if the NEC is respected, wormholes and even throats cannot exist. Furthermore, under the assumptions made, the transformation (5.40) always transfers a flat asymptotic in $\mathbb{M}_\text{E}$ to a flat asymptotic in $\mathbb{M}_\text{J}$ and vice versa (though the corresponding Schwarzschild masses may differ due to nonconstancy of f). If we suppose that there is a twice asymptotically flat wormhole in $\mathbb{M}_\text{J}$, then each of its flat asymptotics has a counterpart in $\mathbb{M}_\text{E}$, and due to smoothness of the transformation we obtain a wormhole in $\mathbb{M}_\text{E}$, contrary to what was said above. Therefore, static asymptotically flat wormholes are absent in $\mathbb{M}_\text{J}$ as well.

This simple reasoning [68, 86] does not require any assumption on the spatial symmetry and works under very general assumptions on the original theory. It admits extensions to dynamic wormholes [86], since in GR they also require NEC violation, although introducing the notion of a throat faces some difficulties in the dynamic case [205]. The asymptotic flatness condition can also be weakened in an evident manner.

It should be stressed that throats in Jordan's picture are not forbidden since local violations of the energy conditions can occur. However, as we have seen, there are no wormholes as global entities.

If the requirements on the function $f(x)$ are weakened, one can obtain wormholes in Jordan's picture in some theories (as we have seen above for a conformal scalar field); however, such wormholes possess some pathological features, for instance, the effective gravitational constant either tends to infinity at the second flat asymptotic [56] or has the wrong sign (see Sec. 5.3.2).

Thus it is rather difficult to obtain more or less realistic wormholes in classical field theory without explicitly invoking phantom fields, at least in GR and its simplest extensions in four dimensions. More opportunities appear in multidimensional theories. As we saw in the previous subsection, in the brane-world scenario the 4D gravitational field equations contain a geometric source which has no reason to respect the usual energy conditions and can thus create wormholes (but there remains the difficulty of solving the full 5D equations). In more general theories of gravity in four dimensions, the terms effectively added to the matter SET as compared with GR can also lead to wormhole formation: an example can be found in the recent paper [226] where a family of static, spherically symmetric wormhole solutions was obtained numerically in dilatonic Einstein-Gauss-Bonnet theory with the Lagrangian

$$L = \text{const} \cdot \left[R + \frac{1}{2}(\partial\phi)^2 + \alpha\, \mathrm{e}^{-\lambda\phi} R^2_{\mathrm{GB}} \right], \tag{5.41}$$

where $R^2_{\mathrm{GB}} = R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma} - 4R_{\mu\nu}R^{\mu\nu} + R^2$ is the Gauss-Bonnet invariant, α and λ are constants. In these solutions, the scalar field is canonical, nonphantom, and there is no other matter violating the NEC, so only the Gauss-Bonnet contribution is responsible for the wormhole's existence. Curiously, among these solutions there are no force-free ones, i.e., those with $g_{00} = \mathrm{e}^{2\gamma} = \text{const}$. Nonexistence of such wormholes has been proved [65] for a class of Lagrangians more general that (5.41), with an arbitrary potential $V(\phi)$ added and an arbitrary function $h(\phi)$ instead of $\mathrm{e}^{-\lambda\phi}$.

Another promising opportunity is connected with quantum phenomena. We have already mentioned the concept of space-time foam existing at Planck scales due to large chaotic quantum fluctuations of the geometry and topology [422]. Among other formations, wormholes are elements of space-time foam. Under usual conditions, the existence of the foam cannot be noticed at macroscopic and even atomic and subatomic scales, but at the earliest stages of the Universe evolution, during its rapid expansion (inflation, see the future chapters), some of the tiny wormholes get a chance to expand to a large size. There exists an idea (see, e.g., [238, 239] and references therein) that such wormholes of kiloparsec size (or like) form a whole chaotic network in the Universe, and the Friedmann homogeneous and isotropic space-time is only obtained as a result of averaging. It has been shown that the inhomogeneities of this foam-like cosmic structure may be observed as manifestations of dark matter (thus solving the well-known problem of mass deficit in galaxies and clusters of galaxies for explaining their dynamic properties [231, 321]). However, the authors themselves admit [240] that by now there are quite a number of unsolved questions in this scenario: what is the wormhole distribution by size and mass? Why are they not yet observed astronomically? Do they take part in the present expansion of the Universe? And so forth. And the same question is probably the most important: which material source supports the existence of large wormholes after they left the Planck scale?

A possible answer is: vacuum polarization due to quantum fields. By quantum field theory, as is well known, the vacuum is the field state with minimum energy and without real particles. This state is not an absolute rest but rather the chaotic creation and annihilation of pairs of all possible virtual particles which all together form some energy density, which is in general not small and can be of any sign. Both depend on the space configuration where the vacuum is considered and on the matter present there. This phenomenon, called vacuum polarization, does really exist and has been repeatedly confirmed by experiment. In particular, the famous Casimir effect, attraction of two closely located plane metal plates in empty space, is explained by a difference in the properties of the electromagnetic field vacuum in the region between the plates and in the whole remaining space.

Calculation of the physical vacuum properties is a difficult task, but some known results look promising from the viewpoint of wormhole existence. Thus, unlike macroscopic matter, the vacuum may not necessarily obey the energy conditions, and it has been found [203, 258] that vacuum polarization can in principle support static wormholes of any size. However, as shown by Hsu [91], the so-called semiclassical matter, suitable for wormhole construction, possesses a special kind of instability due to quantum fluctuations: a wormhole, being left to itself, will be distorted and will be eventually destroyed (but what will then form is not clear). This instability is certainly not universal and not inherent to all wormhole models.

Furthermore: if we discuss on the opportunities of principle, from a physical viewpoint of an "arbitrarily advanced civilization", we can surely admit that, having an interest in interstellar transport, it will launch some negative feedback mechanism that is able to damp the perturbations.

5.4 Observational effects. Wormhole astrophysics

In the recent years, the question of possible wormhole existence in our Universe became a subject of serious analysis, which is evidently caused by the discovery of the Universe's accelerated expansion, repeatedly mentioned in this book. Let us recall the main facts. Acceleration replaced deceleration rather recently from a cosmological viewpoint, about $3–4\cdot 10^9$ years ago and, according to the growing observational evidence, continues to grow. As is well known, if the evolution of an isotropic universe is dominated by a cosmological constant (such that the corresponding density and pressure are related by $\varepsilon + p = 0$), then it develops by a de Sitter law, with a constant acceleration. For the acceleration to grow with time, it is necessary to suppose $\varepsilon + p < 0$, which violates the weak and null energy conditions. (Though, the cosmological constant is also within the observational bounds.) At present, there must be about 70 per cent of this exotic type of matter, called phantom dark energy, and this fraction is increasing. Indeed, unlike ordinary matter, whose density decreases with increasing volume, phantom matter behaves in just the opposite way: its density grows with growing volume.

But is is this kind of exotic matter, whatever be its origin and physical nature, that is required as a construction material for wormholes.

Thus a number of papers discuss the hypothesis that some active galactic nuclei, quasars and/or other compact astrophysical objects can be wormhole entrances, and one seeks specific features able to distinguish wormholes from black holes or other, less familiar objects predicted by theorists (such as boson stars [364], gravastars [292, 303], quasi-black holes [266, 267] etc.) sometimes designated as "black hole mimickers" [390].

Many effects, such as gravitational microlensing and the existence of accretion disks, are common to different classes of compact objects, so it is in general hard to expect that they can reveal specific wormhole features. However, numerous calculations of these effects have very little predictive power, especially if they deal with particular wormhole models. A reason is that there is too great an uncertainty in the nature of the exotic matter supporting the wormhole geometry: as we have seen in Sec. 5.3.2, even under the condition that the scalar curvature is zero, wormhole metrics are obtained with an arbitrary function; in the more general static case there are two arbitrary functions. All this arbitrariness refers to static, spherically symmetric geometries, to say nothing about more general and realistic cases. More than that, if the exotic matter supporting a wormhole occupies a finite volume in space, then the regions outside it are, as always, described by usual, most probably vacuum solutions to the equations of gravity (e.g., Schwarzschild's) with their usual characteristics.

However, there are some effects able to reveal wormhole peculiarities. An interesting point is that a wormhole entrance may be a source of a monopole magnetic field. In such a case, there is no necessity of isolated magnetic poles (monopoles): the magnetic lines of force, as well as electric ones, can simply thread the wormhole, realizing Wheeler's idea of electric or magnetic "charge without charge" [422]). Other distinguishing effects are (see, e.g., [227, 315]) (1) a possible observation of objects in remote parts of the Universe (or another universe) through the throat (assuming that the throat is wide enough and filled with a transparent substance); there is a certain probability to see the same object beside the wormhole and through it with different redshifts [144, 145, 315]; (2) specific modes of particle acceleration and jet shapes due to possible radial magnetic fields; (3) test particle orbits impossible in other settings, such as oscillations across the throat, if there is a minimum of the gravitational potential, (4) gravitational lensing in a strong-field regime [121, 130, 334, 361, 397], etc.

According to Harko et al. [195], a wormhole signature can still be obtained from the properties of accretion disks: it has been obtained that the intensity of a flux emerging from the disk surface is greater for wormholes than for rotating black holes with the same Schwarzschild mass and accretion rate. The conversion efficiency of the accreting mass into radiation was also calculated [195] showning that rotating wormholes provide a much more efficient engine for transformation of the accreting mass into radiation than Kerr black holes.

Curiously, according to [11], it turns out that similar wormholes in the Brans-Dicke theory are "quasi-Schwarzschild objects", and their accretion energy fluxes are smaller by about an order of magnitude as compared to their GR counterparts considered in [195].

Pozanenko and Shatskiy [332] have considered a promising effect connected with observation through a wormhole throat. They remark that if radiation from another universe comes in this way, then a distant observer in our Universe will perceive a wormhole throat like a point source. Furthermore, the brightest known sources are gamma-ray bursts (GRBs). Their brightness in the gamma range may exceed that of the entire sky. If GRB phenomena are present in the other universe (or anywhere beyond the wormhole), an observer in our Universe will detect repeated aperiodic gamma-ray flashes coming from a single point spatially coinciding with the wormhole.

Objects like that do exist: these are sources of soft repeated gamma-ray bursts called soft gamma repeaters. Although most of them can be reliably associated with magnetars in our Galaxy, it still makes sense to check whether or not some of those repeaters can be observational signatures from GRBs of another universe. A calculation of power spectra from GRBs transmitted by a wormhole has led to a conclusion [332] that known soft gamma repeaters are very unlikely to be wormhole candidates.

Thus, for the moment, there are no clear observational traces of natural wormholes in the Universe, but the rapidly growing wealth of observational data leaves a hope that such evidence will sooner or later appear. If, certainly, they really do exist.

This page intentionally left blank

Chapter 6

Stability of spherically symmetric configurations

6.1 Preliminaries

In the previous chapters we have described a number of self-gravitating static and stationary configurations described by solutions to the field equations. To see whether or not such solutions can lead to viable models of some real objects in Nature, one needs to check their stability against various perturbations. This guarantees that the system under consideration will not be immediately destroyed after its formation. The perturbations can be large or small, and can appear as fluctuations in the system itself or be caused by external influence. In all such cases we are dealing with various stability problems.

Perturbations and stability of vacuum and electrovacuum solutions to the Einstein equations (Schwarzschild, Reissner–Nordström, Kerr, Kerr-Newman) have been studied by many authors; a comprehensive review and bibliography can be found in the books [109, 168]. The most important conclusion of these studies is that all these solutions are stable under any small perturbation and have certain spectra of quasinormal modes, i.e., oscillations characterizing their late-time response to any disturbances.

Many papers are devoted to various more general and multidimensional BHs (most often string theory-motivated ones) and their stability, see, e.g., [81, 177] and the recent reviews [213, 256].

Less attention has been paid to solutions with scalar fields although they have some peculiarities of interest. Their main feature is that, unlike gravitational (spin-2) and electromagnetic (spin-1) perturbations, scalar (spin-0) time-dependent perturbations can preserve spherical symmetry. This circumstance is directly related to the Birkhoff theorem which holds for gravitational and electromagnetic fields and in a very rough formulation sounds like that: “spherically symmetric vacuum or electrovacuum space-times are necessarily static” (see more details in Chapter 3), but a scalar field inserts dynamics, so that both the scalar itself and the metric tensor become essentially time-dependent.

Accordingly, in pure vacuum and electrovacuum space-times the multipole expansion of small (linear) perturbations begins with dipole (if there is an electromagnetic field) and quadrupole terms, whereas a scalar field inserts a monopole term. Moreover, it turns out that monopole (spherically symmetric) perturbations are the most “dangerous” in the sense of instability. The point is that in many important cases linear perturbation equations can be reduced to single Schrödinger-like equations with certain effective potentials with respect to some “wave function”, while the role of an energy level is played by a frequency squared. Therefore negative energy levels correspond to imaginary frequencies and hence a possible exponential growth of the perturbations, i.e., instability. The experience shows that higher multipoles lead to positive contributions into the effective potentials, like centrifugal barriers in quantum mechanics. Therefore if a system to be studied is unstable, this instability will most probably manifest itself under monopole perturbations.

In this chapter we will discuss in some detail the stability of static, spherically symmetric solutions to the Einstein equations with scalar fields as sources. Following [67], we will describe a general methodology of studying small spherically symmetric perturbations of scalar-vacuum configurations with any potential $V(\phi)$ (see (6.1)) and pay special attention to space-times with throats. The difficulty with the latter lies in the fact that the effective potentials $V_{\rm eff}(x)$ for perturbations (not to be confused with the self-interaction potential $V(\phi)$) possess a singularity at the throat, which prevents a complete perturbation analysis. It is for this reason that in some previous stability studies of anti-Fisher solutions [14, 58] no unstable modes were found whereas a numerical perturbation analysis of Shinkai and Hayward [370] revealed an instability in the simplest representative of this family of solutions, the Ellis massless wormhole [50, 155].

Gonzalez et al. [175], analyzing the stability of anti-Fisher wormholes, made a proper substitution in the perturbation equation (a special case of the so-called S-deformation [212, 213]) and regularized the effective potential $V_{\rm eff}$. As a result, they were able to consider a previously missed perturbation mode with nonzero perturbation of the throat radius and showed that it can exponentially grow, which means that the anti-Fisher wormholes are unstable. We will verify that a similar methodology can be applied to more general self-gravitating scalar field configurations including those with arbitrary self-interaction potentials $V(\phi)$. To this end, we will prove that, generically, (i) the effective potential $V_{\rm eff}$ for such configurations has precisely the form required for regularization by S-deformation and (ii) any solution of the transformed wave equation with a regularized potential leads to a regular perturbation of the background static configuration.

As a particular example, we study the stability of all anti-Fisher solutions (see Section 4.1) under spherically symmetric perturbations. We prove the instability of Branch A and B solutions and confirm the conclusions of [175] for Branch C solutions (wormholes). A study of particular solutions with nonzero potentials $V(\phi)$ is postponed for the future.

6.2 Perturbation equations

Consider a self-gravitating, minimally coupled scalar field with an arbitrary self-interaction potential in general relativity. The Lagrangian is (up to a constant factor, see, e.g., Section 2.5)

$$L = \sqrt{-g}\Big(R + \epsilon g^{\alpha\beta}\phi_{;\alpha}\phi_{;\beta} - 2V(\phi)\Big), \tag{6.1}$$

where $\epsilon = 1$ for a normal scalar field with positive kinetic energy and $\epsilon = -1$ for a phantom scalar field. Other notations are usual; the gravitational constant is absorbed in the definitions of ϕ and $V(\phi)$. The field equations are

$$\epsilon\Box\phi + V_\phi = 0, \tag{6.2}$$

$$R_\mu^\nu = -\epsilon\phi_{,\mu}\phi^{,\nu} + \delta_\mu^\nu V(\phi), \tag{6.3}$$

where $V_\phi \equiv dV/d\phi$.

The general spherically symmetric metric may be written in the form

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu = e^{2\gamma}dt^2 - e^{2\alpha}du^2 - e^{2\beta}d\Omega^2, \tag{6.4}$$

where α, β, γ are functions of the radial coordinate u and the time coordinate t and $d\Omega^2 = d\theta^2 + \sin^2\theta\, d\varphi^2$. We will also employ the usual notation $r = e^{\beta}$ for the areal radius, such that $4\pi r^2$ is the area of coordinate spheres $u = \text{const}$, $t = \text{const}$. There remains a coordinate freedom in the (u, t) subspace, which in general corresponds to choosing a reference frame preserving the spherical symmetry. In the static case, there is a reference frame such that there is no t-dependence, and then the coordinate freedom concerns the choice of the u coordinate.

We will assume that a certain static, spherically symmetric solution to Eqs. (6.2) and (6.3) is known and study its stability under small spherically symmetric perturbations. We thus consider, instead of $\phi(u)$, a perturbed unknown function

$$\phi(u,t) = \phi(u) + \delta\phi(u,t)$$

and similarly for the metric functions α, β, γ, where $\phi(u)$, etc., are taken from the static solutions.

Preserving only linear terms with respect to time derivatives, we can write all the nonzero component of the Ricci tensor and the time-time component of the Einstein tensor as

$$R_0^0 = e^{-2\gamma}(\ddot{\alpha} + 2\ddot{\beta}) - e^{-2\alpha}[\gamma'' + \gamma'(\gamma' - \alpha' + 2\beta')], \tag{6.5}$$

$$R_1^1 = e^{-2\gamma}\ddot{\alpha} - e^{-2\alpha}[\gamma'' + 2\beta'' + \gamma'^2 + 2\beta'^2 - \alpha'(\gamma' + 2\beta')], \tag{6.6}$$

$$R_2^2 = e^{-2\beta} + e^{-2\gamma}\ddot{\beta} - e^{-2\alpha}[\beta'' + \beta'(\gamma' - \alpha' + 2\beta')], \tag{6.7}$$

$$R_{01} = 2[\dot{\beta}' + \dot{\beta}\beta' - \dot{\alpha}\beta' - \dot{\beta}\gamma'], \tag{6.8}$$

$$G_0^0 = e^{-2\alpha}[2\beta'' + \beta'(3\beta' - 2\alpha')] - e^{-2\beta}, \tag{6.9}$$

where dots and primes denote $\partial/\partial t$ and $\partial/\partial u$, respectively.

6.2.1 General form of the field equations

The zero-order (i.e., static) scalar, $\binom{0}{0}$, $\binom{1}{1}$, $\binom{2}{2}$ components of Eqs. (6.3) and the Einstein equation $G_0^0 = \ldots$ are

$$\phi'' + \phi'(\gamma' + 2\beta' - \alpha') = \epsilon\, e^{2\alpha} V_\phi, \tag{6.10}$$

$$\gamma'' + \gamma'(\gamma' + 2\beta' - \alpha') = -V\, e^{2\alpha}, \tag{6.11}$$

$$\gamma'' + 2\beta'' + \gamma'^2 + 2\beta'^2 - \alpha'(\gamma' + 2\beta') = -\epsilon\phi'^2 - V\, e^{2\alpha}; \tag{6.12}$$

$$-\, e^{2\alpha - 2\beta} + \beta'' + \beta'(\gamma' + 2\beta' - \alpha') = -V\, e^{2\alpha}, \tag{6.13}$$

$$-\, e^{2\alpha - 2\beta} + 2\beta'' + \beta'(3\beta' - 2\alpha') = -\tfrac{1}{2}\epsilon\phi'^2 - V\, e^{2\alpha}. \tag{6.14}$$

The first-order perturbed equations (scalar, $R_{01} = \ldots$, $R_2^2 = \ldots$, and $G_0^0 = \ldots$) read

$$\begin{aligned} &\mathrm{e}^{2\alpha-2\gamma}\delta\ddot{\phi} - \delta\phi'' - \delta\phi'(\gamma' + 2\beta' - \alpha') \\ &\qquad\qquad - \phi'(\delta\gamma' + 2\delta\beta' - \delta\alpha') + \epsilon\delta(\mathrm{e}^{2\alpha}V_\phi) = 0, \end{aligned} \tag{6.15}$$

$$\delta\dot{\beta}' + \beta'\delta\dot{\beta} - \beta'\delta\dot{\alpha} - \gamma'\delta\dot{\beta} = -\tfrac{1}{2}\epsilon\phi'\delta\dot{\phi}, \tag{6.16}$$

$$\begin{aligned} &\delta(\mathrm{e}^{2\alpha-2\beta}) + \mathrm{e}^{2\alpha-2\gamma}\delta\ddot{\beta} - \delta\beta'' - \delta\beta'(\gamma' + 2\beta' - \alpha') \\ &\qquad\qquad - \beta'(\delta\gamma' + 2\delta\beta' - \delta\alpha') = \delta(\mathrm{e}^{2\alpha}V), \end{aligned} \tag{6.17}$$

$$\begin{aligned} &-\delta(\mathrm{e}^{2\alpha-2\beta}) + \delta\beta'' + 6\beta'\delta\beta' - 2\beta'\delta\alpha' - 2\alpha'\delta\beta' \\ &\qquad\qquad = -\epsilon\phi'\delta\phi - \delta(V\,\mathrm{e}^{2\alpha}). \end{aligned} \tag{6.18}$$

Eq. (6.16) may be integrated in t; since we are interested in time-dependent perturbations, we omit the appearing arbitrary function of u describing static perturbations and obtain

$$\delta\beta' + \delta\beta(\beta' - \gamma') - \beta'\delta\alpha = -\tfrac{1}{2}\epsilon\phi'\delta\phi. \tag{6.19}$$

Let us note that we have two independent forms of arbitrariness: one is the freedom of choosing a ***radial coordinate*** u, the other is a ***perturbation gauge***, or, in other words, a reference frame in the perturbed space-time, which can be expressed by imposing a certain relation for $\delta\alpha$, $\delta\beta$, etc. In what follows we will employ both kinds of freedom. All the above equations have been written in the most universal form, without coordinate or gauge fixing.

6.2.2 Gauge $\delta\beta \equiv 0$

This is technically the simplest gauge, in particular, it is convenient for considering usual black hole perturbations, but causes certain difficulties when applied to wormholes and other configurations with throats. The reason is that the assumption $\delta\beta = 0$ leaves the throat radius invariable, while perturbations must in general admit its time dependence [58]. This problem will be discussed below.

With $\delta\beta = 0$, Eq. (6.19) expresses $\delta\alpha$ in terms of $\delta\phi$:

$$2\beta'\delta\alpha = \epsilon\phi'\delta\phi. \tag{6.20}$$

Eq. (6.17) expresses $\delta\gamma' - \delta\alpha'$ in terms of $\delta\alpha$ and $\delta\phi$:

$$\beta'(\delta\gamma' - \delta\alpha') = 2\,\mathrm{e}^{2\alpha-2\beta}\delta\alpha - \delta(\mathrm{e}^{2\alpha}V). \tag{6.21}$$

Substituting all this into (6.15), we obtain the wave equation

$$e^{2\alpha-2\gamma}\delta\ddot{\phi} - \delta\phi'' - \delta\phi'(\gamma' + 2\beta' - \alpha') + U\delta\phi = 0,$$
$$U \equiv e^{2\alpha}\left\{\epsilon(V - e^{-2\beta})\frac{\phi'^2}{\beta'^2} + \frac{2\phi'}{\beta'}V_\phi + \epsilon V_{\phi\phi}\right\}. \tag{6.22}$$

Before proceeding with a study of the wave equation, let us make sure that all the remaining Einstein equations hold as a consequence of (6.20), (6.15) and (6.17) and do not lead to any new restrictions. Consider the component (6.18) (the constraint equation). It now takes the form

$$2\delta\alpha\, e^{2\alpha-2\beta} + 2\beta'\delta\alpha' = \epsilon\phi'\delta\phi + \delta(V\, e^{2\alpha}). \tag{6.23}$$

This equation holds automatically owing to the zero-order equations and (6.20). Indeed, a substitution of $\delta\alpha$ from (6.20) brings (6.23) to the form

$$\frac{\delta\phi}{\beta'}\left[\beta'\phi'' - \phi'\beta'' + e^{2\alpha-2\beta}\phi' - e^{2\alpha}V\phi' - \epsilon\, e^{2\alpha}V_\phi\beta'\right] = 0, \tag{6.24}$$

Now, substituting ϕ'' and β'' from (6.10) and (6.13), respectively, we see that all terms cancel, i.e., the equation does hold. Furthermore, the Einstein equation $G_1^1 = \ldots$ holds as a consequence of (6.17) and (6.18); lastly, the equation $G_2^2 = \ldots$ holds due to the Bianchi identity $\nabla_\nu G_1^\nu = 0$ and the corresponding property of the stress-energy tensor of the scalar field.

So we can return to Eq. (6.22). Passing on to the "tortoise" coordinate x introduced according to

$$du/dx = e^{\gamma-\alpha}, \tag{6.25}$$

and changing the unknown function $\delta\phi \mapsto \psi$ according to

$$\delta\phi = \psi(x,t)\, e^{-\beta}, \quad \Leftrightarrow \quad \psi(x,t) = r\delta\phi, \tag{6.26}$$

we reduce the wave equation to its canonical form, also called the master equation for radial perturbations:

$$\ddot{\psi} - \psi_{xx} + V_{\text{eff}}(x)\psi = 0, \tag{6.27}$$

(the index x denotes d/dx), with the effective potential

$$V_{\text{eff}}(x) = e^{2\gamma-2\alpha}[U + \beta'' + \beta'^2 + \beta'(\gamma' - \alpha')]. \tag{6.28}$$

This effective potential was previously obtained in other notations for $\epsilon = -1$ in [126]. A further substitution

$$\psi(x,t) = y(x)\, e^{i\omega t}, \quad \omega = \text{const}, \tag{6.29}$$

which is possible because the background is static, leads to the Schrödinger-like equation

$$y_{xx} + [\omega^2 - V_{\text{eff}}(x)]y = 0. \tag{6.30}$$

If there is a nontrivial solution to (6.30) with $\mathrm{Im}(\omega) < 0$ satisfying some physically reasonable conditions at the ends of the range of u (in particular, the absence of ingoing waves that guarantees the absence of energy pumping into the system from outside), then the static system is unstable since $\delta\phi$ can grow exponentially with t. Otherwise our static system is stable in the linear approximation. Thus, as usual in such studies, the stability problem is reduced to a boundary-value problem for Eq. (6.30) — see, e.g., [58, 70–72, 109, 175, 254, 255].

Note that all the above relations are written without fixing the background radial coordinate u.

6.2.3 Gauge-invariant perturbations

To be sure that we are dealing with real perturbations of the static background rather than purely coordinate effects, it is necessary to construct gauge-invariant quantities.

Small coordinate transformations $x^a \mapsto x^a + \xi^a$ in the (t, u) subspace can be written as

$$t = \overline{t} + \Delta t(t, u), \qquad u = \overline{u} + \Delta u(t, u), \tag{6.31}$$

where Δt and Δu are supposed to be small. Any scalar quantity with respect to such transformations, such as, e.g., $\phi(t, u)$ acquires an increment:

$$\Delta\phi = \dot{\phi}\Delta t + \phi'\Delta u \approx \phi'\Delta u, \tag{6.32}$$

in the linear approximation since both $\dot{\phi}$ and Δt are small. The quantity r, also being a scalar in the (t, u) subspace (a 2-scalar, for short), behaves in the same way. If there are perturbations $\delta\phi$ and δr, the transformation (6.31) changes them as follows:

$$\begin{aligned} \delta\phi \mapsto \overline{\delta\phi} &= \delta\phi + \phi'\Delta u, \\ \delta r \mapsto \overline{\delta r} &= \delta r + r'\Delta u. \end{aligned} \tag{6.33}$$

It then follows that the combination

$$\psi_1 \equiv r'\delta\phi - \phi'\delta r, \tag{6.34}$$

is invariant under the transformation (6.31), or gauge-invariant. Recall that the prime here denotes d/du in the background static configuration.

One can notice that combinations constructed like (6.34) from any 2-scalars (for example, e^{ϕ} and $\beta = \ln r$, or two different linear combinations of ϕ and r) are also gauge-invariant. Moreover, ψ_1 multiplied by any 2-scalar or any combination of background quantities which are known and fixed functions of u is gauge invariant.

The physical properties of perturbations must not depend on which gauge-invariant quantity ψ is chosen to describe them. Meanwhile, with different ψ, the effective potentials will, in general, also be different. However, given a specific background configuration, in order that the theory be consistent, these different potentials should lead to the same perturbation spectrum.

Due to gauge invariance of ψ_1, equations that govern it may be written in any admissible gauge. In particular, we can choose the gauge $\delta\beta = 0$, and Eq. (6.27) for ψ may then be considered as a result of substituting $\psi = (r/r')\psi_1 = r\delta\phi$ in a manifestly correct equation for $\psi_1 = r'\delta\phi$.

On the similar problem regarding cosmological perturbations and the definition of the corresponding gauge invariants, see, e.g., the review [47].

6.2.4 Regularized potential near a throat

The gauge $\delta\beta = 0$ (the same as $\delta r = 0$) is suitable for describing the perturbations at any points except those with $r' = 0$: these are throats and other critical points of $r(u)$. Indeed [58], putting $\delta r = 0$, we forbid perturbations of the throat radius, while there is no physical reason for that. Technically, this restriction manifests itself in a generically infinite value of the potential $U(u)$ in Eq. (6.22) and consequently in $V_{\rm eff}$ involved in the wave equation (6.27). Throats are only possible in the case $\epsilon = -1$ and, provided $U/r^2 < 1$ at such a throat ($u = u_{\rm th}$), the potential there is a wall of infinite height, with the generic behavior $V_{\rm eff} \sim 1/(u - u_{\rm th})^2$ near the throat since we have there generically $r'(u) \sim u - u_{\rm th}$. As a result, perturbations that are independent at different sides of the throat necessarily turn to zero at the throat itself, and we thus partly lose information on their possible properties. Such an incomplete treatment has led to a conclusion that anti-Fisher wormholes [50, 155] (that is, wormhole solutions to Eqs. (6.2), (6.3) with $V \equiv 0$ and $\epsilon = -1$) were stable under spherically symmetric perturbations. A similar conclusion was made in [58] concerning such wormholes and cold black holes by using another (harmonic) gauge, $\delta\alpha = 2\delta\beta + \delta\gamma$, which does not lead to a pole in the effective potential, but as follows from our further consideration in this chapter, this analysis was also incomplete.

It could seem that the above difficulty only concerns the gauge $\delta\beta = 0$. However, due to the gauge-invariant nature of Eq. (6.27) (to be verified below), it is clear that the problem is inherent to the background geometry itself, and the pole in the effective potential always emerges at a throat, if any.

A way of avoiding the restriction $\delta\beta(u_{\rm th}) = 0$ is connected with the so-called S-deformations of the potential $V_{\rm eff}$. This method was used in [212, 213] for transforming a partly negative potential to a positive-definite one in master equations for perturbations of higher-dimensional black holes. Using this method, Gonzalez *et al*. [175] transformed a singular potential to a nonsingular one for perturbations of the anti-Fisher wormholes and discovered the existence of an exponentially growing mode, showing that such wormholes are unstable. Let us try to formulate a similar scheme suitable for the more general field system (6.1).

Consider a wave equation of the type (6.27)

$$\ddot{\psi} - \psi_{xx} + W(x)\psi = 0, \tag{6.35}$$

with an arbitrary potential $W(x)$ (whose specific example is the above potential $V_{\rm eff}$). If there is a function $S(x)$ such that $W(x)$ is presented in the form

$$W(x) = S^2(x) + S_x, \tag{6.36}$$

then Eq. (6.35) is rewritten as follows:

$$\ddot{\psi} + (\partial_x + S)(-\partial_x + S)\psi = 0. \tag{6.37}$$

Now, if we introduce the new function

$$\chi = (-\partial_x + S)\psi, \tag{6.38}$$

then, applying the operator $-\partial_x + S$ to the left-hand side of Eq. (6.37), we obtain the following wave equation for χ:

$$\ddot{\chi} - \chi_{xx} + W_{\rm reg}(x)\chi = 0, \tag{6.39}$$

with the new effective potential

$$W_{\rm reg}(x) = -S_x + S^2 = -W(x) + 2S^2. \tag{6.40}$$

If a static solution $\psi_s(x)$ of Eq. (6.35) is known, so that $\psi_{s,xx} = W(x)\psi_s$, then we can choose

$$S(x) = \psi_{s,x}/\psi_s, \tag{6.41}$$

to carry out the above transformation.

Generically, the function U in (6.22) and hence the potential (6.28) behave near a throat as $r'^{-2} \sim (u - u_{\rm th})^{-2} \sim x^{-2}$, where, without loss of generality, we put $x = 0$ at the throat. Assuming that the potential $W(x)$ behaves in such a way, let us look if a transition to $W_{\rm reg}$ can really remove this singularity. Above all, we see that according to (6.40), such removal is only possible if $W \to +\infty$ as $x \to 0$ since we are dealing with real quantities. Thus a potential wall in $W(x)$ can be removed but a potential well cannot.

A positive pole $W \sim x^{-2}$ can be removed in $W_{\rm reg}$ if

$$S \approx -1/x \quad \Rightarrow \quad \psi_s \propto 1/x. \tag{6.42}$$

It is a necessary condition for regularizing the potential. Besides, to avoid a singularity of $W_{\rm reg}$, ψ_s must be nonzero in the whole range of u. Moreover, according to (6.40), it is clear that near the throat $x = 0$ in this case

$$W(x) \approx 2/x^2. \tag{6.43}$$

Thus, to be regularized by the procedure described, the pole in $W(x)$ must behave as (6.43). Let us show that this is generically the case for the potential $W = V_{\rm eff}(x)$ given by Eq. (6.28).

Suppose such a generic situation, so that

(i) the function $\beta(x)$ in the background metric is expanded near its minimum (the throat) in powers of x as follows:

$$\beta(x) = \beta_0 + \tfrac{1}{2}\beta_2 x^2 + \tfrac{1}{6}\beta_3 x^3 + \ldots . \tag{6.44}$$

where $\beta_{0,2,3}$ are constants;

(ii) the background quantity $\phi'(x) \neq 0$ at $x = 0$.

Here and till the end of the section, we use the coordinate freedom to choose the "tortoise" radial coordinate $u = x$ specified by the condition $\alpha = \gamma$. All functions are considered as power series in x at small x.

Let us estimate $V_{\rm eff}$ (6.28). The term that determines the pole at $x = 0$ is

$$W_{\rm pole}(x) = \mathrm{e}^{2\gamma}(V - \mathrm{e}^{-2\beta})\frac{\phi'^2}{\beta'^2}, \tag{6.45}$$

where we have put $\epsilon = -1$ since throats are possible only in this case. By (6.44),

$$\beta_x = \beta_2 x + \tfrac{1}{2}\beta_3 x^2 + \ldots . \tag{6.46}$$

Now we use the equations governing the static configuration: from (6.13) it follows that $\mathrm{e}^{2\gamma}(\mathrm{e}^{-2\beta} - V) = \beta_{xx}$ at $x = 0$, and then from (6.14) we find that $\phi_x^2 = 2\beta_{xx}$ at $x = 0$. Thus from assumption (ii) it follows that $\beta_2 \neq 0$. Substituting all this to (6.45), we find that it behaves precisely as required in (6.43).

Thus we have shown that, for a (generic) throat in a solution to Eqs. (6.10)–(6.14), the effective potential $V_{\rm eff}$ for spherically symmetric perturbations satisfies the necessary condition for regularization by the above method.

Whether or not this regularization really works and leads to a regular boundary-value problem for the perturbations should be investigated for specific background configurations. (In particular, one should also take into account the other singular term in the potential $V_{\rm eff}$, proportional to V_ϕ/β_x: the terms $\propto 1/x$ evidently depend on the finite part of $S(x)$.)

A positive example of such a study, concerning anti-Fisher wormholes, is known from [175]; we will show that the other two branches of the anti-Fisher solution are also such examples.

6.2.5 Regular perturbations near a throat

Suppose we have found a solution $\chi(x,t)$ to Eq. (6.39), satisfying the appropriate boundary conditions. The function χ is regular at $x=0$ since the potential $W_{\rm reg}$ is regular there. If χ is a growing function of t, it probably indicates an instability of the initial static configuration; but this is indeed the case only if $\chi(x,t)$ creates regular perturbations of the metric functions α, β, γ and the scalar field ϕ.

Let us look at how it happens. Given $\chi(x,t)$, a solution to (6.35), or (6.27), is found as $\psi = (\partial_x + S)\chi$. Generically, χ is finite at $x=0$ while $S \approx 1/x$ at small x, hence $\psi \sim 1/x$, and according to (6.26) we obtain $\delta\phi \to \infty$ at the throat. This result is in fact quite natural since the relation (6.26) corresponds to the gauge $\delta r = 0$, in which the throat radius is fixed, whereas we are seeking perturbations with nonzero δr on the throat. So it is necessary to pass on to another gauge, which is easily done due to gauge invariance of the quantity ψ given by

$$\psi = r\delta\phi - \frac{r\phi_x}{r_x}\delta r. \tag{6.47}$$

Namely, a finite expression for δr is obtained in the gauge $\delta\phi = 0$ provided $\phi_x(0) \neq 0$ since then

$$\delta r = -\frac{r_x}{r\phi_x}\psi, \tag{6.48}$$

while the product $r_x\psi$ is finite. It remains to find $\delta\alpha$ and $\delta\gamma$ from the perturbation equations in the gauge $\delta\phi = 0$. From Eqs. (6.16) and (6.15) we find

$$\beta_x\delta\alpha = \delta\beta_x + \delta\beta(\beta_x - \gamma_x), \tag{6.49}$$

$$\delta\gamma_x = \delta\alpha_x - 2\delta\beta_x - \frac{2\epsilon}{\phi_x}V_\phi\, e^{2\alpha}\delta\alpha, \tag{6.50}$$

but here we are again facing a problem: according to (6.49), in general $\delta\alpha$ diverges at the throat where $\beta_x = 0$. This divergence is only avoided if the right-hand side of (6.49) behaves like $\beta_x \sim u - u_{\rm th} \sim x$.

Surprisingly, this is the case in a generic situation, as can be verified in a general form using near-throat expansions. Indeed, let us preserve the above assumptions (i) and (ii) and assume, in addition, that

(iii) the function $\chi(x,t)$ that solves Eq. (6.39) is finite and nonzero at $x = 0$,
(iv) the function $S(x)$ behaves at small x according to (6.42).

Our task is to estimate the right-hand side of Eq. (6.49) in the order $O(x^0)$: if it is zero, it means that $\beta_x \delta\alpha \sim x \sim \beta_x$ and thus $\delta\alpha(0)$ is finite; the remaining metric perturbation $\delta\gamma$ is then found from (6.50) and is also finite.

Taking $\delta\beta = \delta r/r$ from (6.48) and substituting ψ as $\psi = \chi_x + S\chi \sim -\chi_0/x$, where $\chi_0 = \chi(0)$, we find

$$\delta\beta = \frac{\chi_0\, \mathrm{e}^{-\beta}}{\phi_x}\Big(\beta_2 + \frac{1}{2}\beta_3 x + \ldots\Big).$$

Then we substitute this expression to (6.49) to obtain

$$\beta_x \delta\alpha \approx \frac{\chi_0\, \mathrm{e}^{-\beta}}{\phi_x^2}\Big[-\phi_{xx}\beta_2 + \frac{1}{2}\phi_x(\beta_3 - 2\gamma_x\beta_2)\Big],$$

where all quantities are taken at $x = 0$. Now, ϕ_{xx} can be expressed from the background equation (6.10), β_{xx} from (6.13); we can use the fact that $\beta_2 = \beta_{xx}(0)$ etc., and we can also ignore all terms proportional to β_x. After these substitutions we finally obtain

$$\beta_x \delta\alpha \propto \mathrm{e}^{2\gamma} V_\phi \Big[\epsilon\beta_{xx} + \frac{1}{2}\phi_x^2\Big]_{x=0}.$$

But the expression in the square brackets vanishes due to the difference of Eqs. (6.11) and (6.12), which proves that $\beta_x\delta\alpha = O(x)$ and thus $\delta\alpha(0)$ is finite.

We conclude that *under the generic assumptions (i)–(iv), regularization of the potential V_{eff} always leads to finite perturbations of the background static solution.*

6.3 Instabilities of the Fisher and anti-Fisher solutions

6.3.1 The static solutions

Let us recall the well-known static, spherically symmetric solutions to the field equations (6.2) and (6.3) for zero potential, $V \equiv 0$. For $\varepsilon = +1$ this corresponds to Fisher's solution [164] and for $\varepsilon = -1$ to that of Bergmann and Leipnik [35] (but these authors used the curvature coordinates [i.e., the condition $u \equiv r$ in terms of the metric (6.4)], which are not well suited for the problem, and maybe therefore they did not give a clear interpretation of the solution). These solutions are described in detail in Sec. 4.1.4 and 5.3.2, but here we will briefly recall them using slightly different notations.

The solution can be written jointly for $\epsilon = \pm 1$ using the harmonic coordinate $u = v$ in the metric (6.4), corresponding to the coordinate condition $\alpha(v) = \gamma(v) + 2\beta(v)$ [50]. The solution reads

$$ds^2 = \mathrm{e}^{-2mv} dt^2 - \frac{\mathrm{e}^{2mv}}{s^2(k,v)}\left[\frac{dv^2}{s^2(k,v)} + d\Omega^2\right], \qquad \phi = Cv, \tag{6.51}$$

where the integration constants m (the Schwarzschild mass), C (the scalar charge) and k are related by the equality

$$2k^2 \operatorname{sign} k = 2m^2 + \epsilon C^2. \tag{6.52}$$

The function $s(k,v)$ is defined according to (4.18). The coordinate v is defined in the whole range $v > 0$ for $k \geq 0$ and in the range $0 < v < \pi/|k|$ for $k < 0$. The value $v = 0$ in all cases corresponds to flat spatial infinity: at small v, the spherical radius is $r(v) \approx 1/v$, and the metric becomes approximately Schwarzschild with mass m.

In the case $k > 0$, we pass over to the quasiglobal coordinate u (such that $\alpha + \gamma = 0$ in (6.4)) by the transformation

$$\mathrm{e}^{-2kv} = 1 - 2k/u =: P(u), \tag{6.53}$$

and the solution takes the form

$$ds^2 = P^a dt^2 - P^{-a} du^2 - P^{1-a} u^2 d\Omega^2, \quad \phi = -\frac{C}{2k} \ln P(u), \tag{6.54}$$

with the constants related by

$$a = m/k, \quad a^2 = 1 - \epsilon C^2/(2k^2). \tag{6.55}$$

The Fisher solution [164] corresponds to $\epsilon = +1$, it consists of a single branch $k > 0$ and, in (6.54), $|a| < 1$. It is defined in the range $u > 2k$, and $u = 2k$ is a naked central ($r = 0$) singularity which is attractive for $m > 0$ and repulsive for $m < 0$. The Schwarzschild solution is restored at $C = 0$, $a = 1$ for $m > 0$ and at $C = 0$, $a = -1$ for $m < 0$.

The solution for $\epsilon = -1$ (the anti-Fisher solution) splits into three branches with the following basic properties.

Branch A, $k > 0$: we have again the solution (6.54), but now $|a| > 1$. For $m < 0$, that is, $a < -1$, we have, as in the Fisher solution, a naked central singularity at $u = 2k$. The situation is, however, drastically different for $m > 0$, $a > 1$: there is a throat (a regular minimum of r) at $u = u_{\rm th} = (a+1)k$, and $r \to \infty$ as $u \to 2k$. Moreover, for $a = 2,\ 3,\ \ldots$ the metric exhibits a horizon of order a at $u = 2k$ and admits a continuation to smaller u [57]. These horizons have an infinite area and zero Hawking temperature (the so-called cold black holes [58]). The throat radius does not coincide with the (geometrized) Schwarzschild mass $m = ak$ but is of the same order of magnitude.

The metric (6.54) has a curvature singularity at $u = 2k$ if $a < 2$ (except for the Schwarzschild case $a = 1$), a finite curvature if $a = 2$ and zero curvature if $a > 2$. At noninteger $a > 2$, we are dealing with weaker singularities termed "singular horizons" (see Sec. 4.1 and [58]).

Branch B, $k = 0$: the solution is defined in the range $u \in \mathbb{R}_+$ and is rewritten in terms of the quasiglobal coordinate $u = 1/v$ as follows:

$$ds^2 = \mathrm{e}^{-2m/u} dt^2 - \mathrm{e}^{2m/u}[du^2 + u^2 d\Omega^2], \quad \phi = C/u. \tag{6.56}$$

As before, $u = \infty$ is a flat infinity, while at the other extreme, $u \to 0$, the behavior is different for positive and negative mass. For $m < 0$, $u = 0$ is a singular centre ($r = 0$), while for $m > 0$, $r \to \infty$, and all $K_i \to 0$ as $u \to 0$. This is again a singular horizon: the nonanalyticity of the metric in terms of u makes its continuation impossible. The throat occurs at $u = m$ and has the size $e \cdot m$, e being the base of natural logarithms.

Branch C, $k < 0$: the solution describes a wormhole with two flat asymptotics at $v = 0$ and $v = \pi/|k|$. The metric has the form [50, 155]

$$\begin{aligned} ds^2 &= \mathrm{e}^{-2mv} dt^2 - \frac{k^2\,\mathrm{e}^{2mv}}{\sin^2(kv)}\left[\frac{k^2\,du^2}{\sin^2(kv)} + d\Omega^2\right] \\ &= \mathrm{e}^{-2mv} dt^2 - \mathrm{e}^{2mv}[du^2 + (k^2 + u^2) d\Omega^2], \end{aligned} \tag{6.57}$$

where v is expressed in terms of the quasiglobal coordinate u, defined on the whole real axis, by $\overline{k}v = \cot^{-1}(u/\overline{k})$, where we denote $-k = \overline{k} > 0$. If $m > 0$, the wormhole is attractive for ambient test matter at the first asymptotic ($u \to \infty$) and repulsive at the second one ($u \to -\infty$), and vice versa in case $m < 0$. The wormhole throat occurs at $u = m$.

6.3.2 Perturbations: The Fisher solution

For a massless scalar field [$V \equiv 0$ in (6.1)], the effective potential in (6.27) takes the following form in terms of the quasiglobal coordinate u:

$$V_{\rm eff} = -\epsilon \frac{A\phi'^2}{r'^2} + \frac{Ar''}{r} + \frac{A'r'}{r} = -\epsilon \frac{A\phi'^2}{r'^2} + \frac{A}{r^2} - \frac{A^2 r'^2}{r^2}, \tag{6.58}$$

where $A(u) := e^{2\gamma(u)} = e^{-2\alpha(u)}$, and the second equality in (6.58) follows from Eq. (6.13).

Calculating $V_{\rm eff}$ for the solution (6.54), we find a common expression for both $\epsilon = +1$ and $\epsilon = -1$:

$$V_{\rm eff}(u) = kP^{2a} \frac{2au^3 - 3(1+a)^2 ku^2 + 2(3+4a+3a^2+2a^3)k^2 u - (1+a)^4 k^3}{u^2(u-k(1+a))^2(u-2k)^2}. \tag{6.59}$$

Since in the Fisher solution $a < 1$, the binomial $u - k(1+a)$ is positive at all $u \geq 2k$, and the only singularity in $V_{\rm eff}$ is $u \to 2k$, coinciding with the singularity of the background solution. Near the singularity, at which according to (6.25) we can put $x = 0$, $V_{\rm eff}(u) \sim -1/(4x^2)$, a negative pole in agreement with [72] and many subsequent papers.

The boundary condition at spatial infinity ($u \to \infty$, $x \to \infty$) is natural: $\delta\phi \to 0$, or $\psi \to 0$. For $u \to 2k$, where the background field ϕ tends to infinity, the boundary condition is not so evident. In [72] and other papers, dealing with minimally coupled or dilatonic scalar fields, the minimal requirement was used providing the validity of the perturbation scheme:

$$|\delta\phi/\phi| < \infty. \tag{6.60}$$

(The requirement of absence of ingoing waves then does not lead to further restrictions.) Under this boundary condition it is easy to conclude that there are solutions to the Schrödinger-like equation (6.30) with any $\omega^2 < 0$, which means that the static field configuration is unstable, in agreement with the previous work [72] (see also [57] for details).

6.3.3 Perturbations: The anti-Fisher solution

Branch A

The effective potential has the same form (6.59), but now, since $a > 1$ (we restrict ourselves to this case providing $m > 0$), the potential has a positive pole at $u = u_{\rm th} = k(a+1) > 2k$, the throat in the background configuration.

To remove this singularity, we were able to find the following simple static solution to Eq. (6.27):

$$\psi_{s+}(u) \propto r(u)\frac{au - u_{\rm th}}{u - u_{\rm th}}. \tag{6.61}$$

Applying the technique described in Sec. 6.2.4 with these $\psi_s(u)$, we calculate the new effective potential $W_{\rm reg} = W_A(u)$:

$$\begin{aligned} W_A(u) &= \left(1 - \frac{2k}{u}\right)^{2a} \frac{N(u)}{u^2(u-2k)^2(au-u_{\rm th})^2}, \\ N(u) &= 3(1+a)^4k^4 - 2a(6+19a+16a^2+3a^3)k^3u \\ &\quad +3a^2(9+10a+a^2)k^2u^2 - 2a^2(4+5a)ku^3 + 2a^2u^4. \end{aligned} \tag{6.62}$$

The potential (6.62) has no singularities at $r > 2k$ and is partly negative, which should in general lead to an instability. To prove it, one can use the method of time-domain integration [186] allowing for the time evolution of the perturbations under prescribed initial and boundary conditions. The latter, as usual, must provide the absence of ingoing waves from the boundary, and in our case it is sufficient to simply require $\psi \to 0$ and $\chi \to 0$ as $x \to \pm\infty$ for all three branches of the anti-Fisher solution. The reason is that the effective potential $V_{\rm eff}$ as well as the regularized potentials vanish at large $|x|$, and therefore those modes of interest with $\omega^2 < 0$ should exponentially decay at large $|x|$.

Examples of plots for the potential (6.62) and the results of time-domain integration are shown in Fig. 6.1. By fitting of the profile we find that the perturbations grow approximately as $\psi \propto e^{0.25t/m}$.

In the limit $a \to 1$ the regularized potential still has a negative gap; however, at $a = 1$ no growing mode is observed, and a stationary solution dominates at late times (Fig. 6.2), while at any nonzero $a - 1$ the perturbation does grow. This shows how the instability is "dying out" when approaching the Schwarzschild solution. One can recall that in the genuine Schwarzschild case there is no scalar field, and the modes we are considering here simply do not exist.

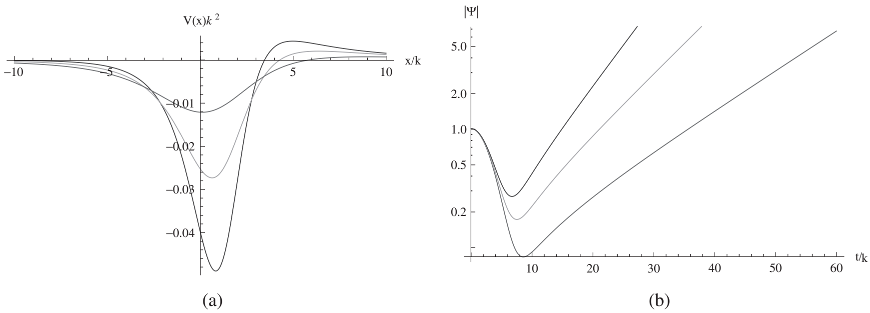


Figure 6.1 The regularized effective potential W_A (left panel) and the time-domain profiles (right panel) for the Branch A solutions with $a = 3/2$ (blue), $a = 2$ (green), $a = 3$ (red). Smaller values of a correspond to deeper potential wells and a more rapid growth of perturbations.

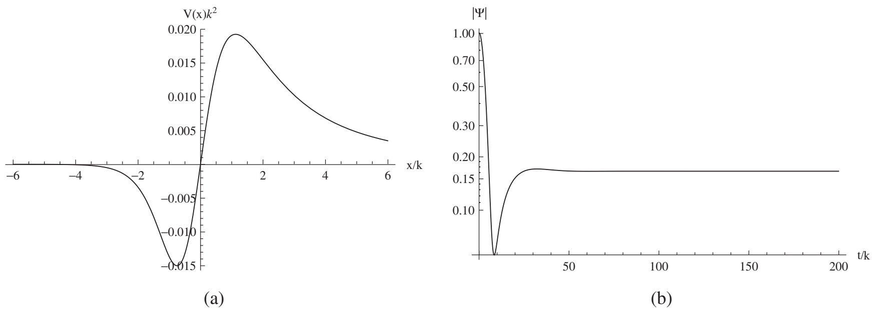


Figure 6.2 The potential W_A (left panel) and the time-domain profile (right panel) for the Branch A solution with $a = 1$.

Branch B

For the solution (6.56), the effective potential (6.58)

$$V_{\text{eff}}(u) = m \exp\left(-\frac{4m}{u}\right) \frac{2u^3 - 3mu^2 + 4m^2u - m^3}{(u-m)^2 u^4}, \tag{6.63}$$

is singular at $u_{\text{th}} = m$. We again find a static solution to (6.27)

$$\psi_{s0}(u) \propto \frac{u^2 \, \mathrm{e}^{m/u}}{u - m}, \tag{6.64}$$

and perform the transformation described in Sec. 2.5.

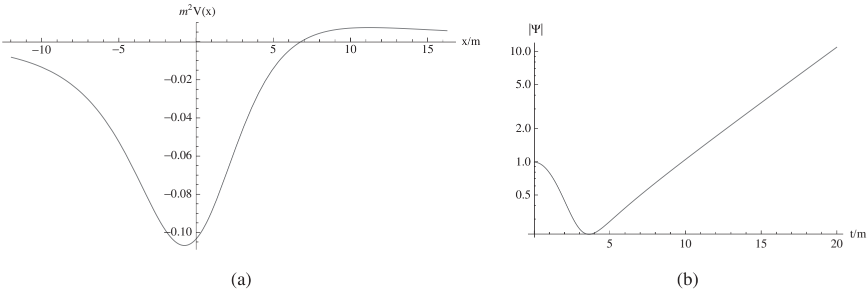


Figure 6.3 The potential W_B (left panel) and the time-domain profile (right panel) for the Branch B solution.

The regularized effective potential has the form of an inverse potential barrier:

$$W_B(u) = \exp\left(-\frac{4m}{u}\right)\frac{3m^2 - 10mu + 2u^2}{u^4}. \tag{6.65}$$

The perturbations grow approximately as $\psi \propto e^{0.23t/m}$ (see example in Fig. 6.3).

Branch C

For the wormhole solution (6.57), the potential (6.58) has the form

$$V_{\text{eff}}(u) = \mathrm{e}^{-4mv}\frac{2k^4 + k^2(3m^2-2mu+3u^2) - m(m^3-4m^2u+3mu^2-2u^3)}{(m-u)^2(k^2+u^2)^2}. \tag{6.66}$$

It again exhibits a positive pole on the throat. To regularize it, we can take any of the two static solutions

$$\psi_{s-1}(u) \propto \frac{r(u)}{u-m}[k^2 + m^2 - mv(k^2 + mu)], \tag{6.67}$$

$$\psi_{s-2}(u) \propto \frac{r(u)}{u-m}(k^2 + mu). \tag{6.68}$$

The first of them coincides with the solution found by González ***et al***. in [175]. As a result, we reproduce their regular potential

$$W_{C1}(u) = e^{-4mv}\frac{N_1(u)}{(k^2+u^2)^2(k^2(mv-1)+m^2(uv-1))^2}, \tag{6.69}$$

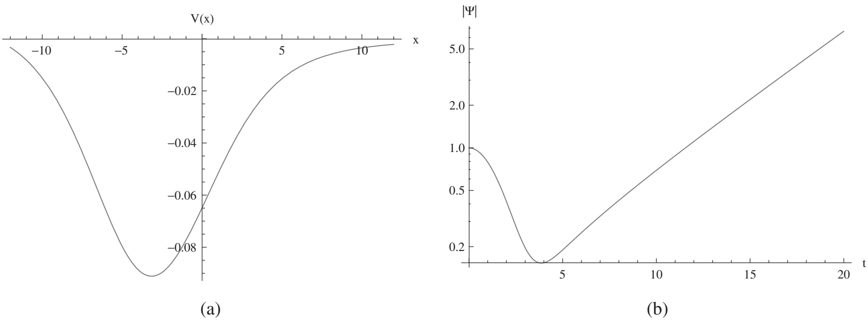


Figure 6.4 The potential W_{C1} (left panel) and the time-domain profile (right panel) for Branch C ($k = m = 1$).

where

$$N_1(u) = -3k^6(mv-1)^2 + m^4(3m^2 - 10mu + 2u^2)(uv-1)^2 + k^2m^3(uv-1)(8u + 2m^2v + m(5 - 15uv)) + k^4m(mv-1)(2u + m^2v + m(9 - 12uv)).$$

A further investigation revealing an unstable mode of perturbation is described in detail in [175] (Fig. 6.4).

The second static solution (6.68) can be used to demonstrate that, despite the different regularized potentials, the physical result, namely, the perturbation growth rate remains the same. Indeed, the regularized potential produced by the solution (6.68) reads

$$W_{C2}(u) = e^{-4mv}\frac{N_2(u)}{(k^2+mu)^2(k^2+u^2)^2},$$
$$N_2(u) = -3k^6 + k^4m(m-12u) + k^2m^2(2m-15u)u + m^2u^2(3m^2 - 10mu + 2u^2), \quad (6.70)$$

and leads to perturbations growing as shown in Fig. 6.5, at the same rate as in Fig. 6.4.

We have demonstrated the instability of all anti-Fisher solutions to the Einstein-scalar equations under spherically symmetric perturbations, thus confirming and extending the conclusions of [175] made for Branch C solutions (wormholes). It turns out that in almost all cases the characteristic time of perturbation growth is of the order of the time needed for a light signal to cover a distance equal to the throat radius.

We have found out that the S-deformation method of regularizing the effective potential for the perturbations, having a positive pole at a throat

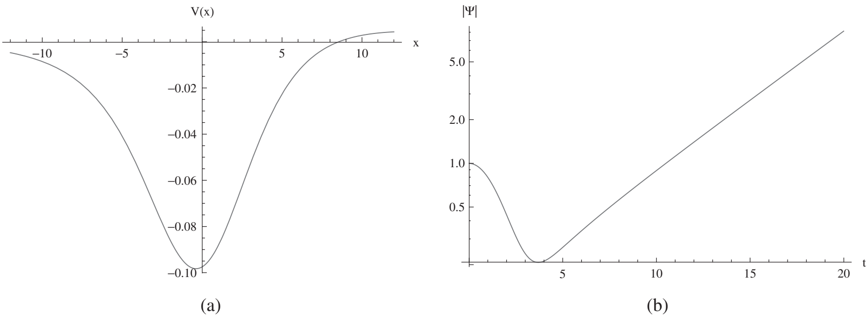


Figure 6.5 The potential W_{C2} (left panel) and the time-domain profile (right panel) for Branch C ($k = m = 1$).

(a minimum of the spherical radius $r(u)$), used in [175], is applicable to any static, spherically symmetric configuration of self-gravitating scalar fields with self-interaction potentials. We have shown in a general form that, under some generic assumptions [items (i)–(iv) in Sec. 6.2], the potential V_{eff} near the throat has the form admitting regularization, and that a regular mode found as a solution to the regularized equation leads to a regular perturbation of the initial configuration. The latter circumstance is quite nontrivial because regularity is required not only for solutions to the master equation but also for all metric coefficients.

However, application of this methodology to specific solutions with self-interaction potentials $V(\phi) \neq 0$ faces the problem of explicitly finding the function $S(x)$ satisfying Eq. (6.36), or, equivalently, a proper static solution to the master equation (a zero mode). We hope to extend this study to some configurations with nonzero $V(\phi)$ (including black universe solutions desribed above) in the near future.

6.4 Extensions and related problems

An important extension of the presently obtained results is connected with the fact that Eqs. (6.2), (6.3) are not only the equations of GR but also the Einstein-frame equations of the general scalar-tensor (STT) (see Sec. 4.1) and $f(R)$ theories of gravity [284, 294, 309, 412]. Note that in $f(R)$ theories, the Einstein frame always contains a scalar field with nonzero $V(\phi)$. The observable physics in such theories is usually described in the corresponding Jordan conformal frame, whose metric differs from that of the

Einstein frame only by a conformal factor that varies from theory to theory. A transition from one frame to the other is, from the viewpoint of differential equations, simply a substitution, therefore the stability study can be safely performed in the Einstein frame. Then, if the conformal factor is everywhere regular and nonzero (including the end points of the coordinate range), the conformal mapping preserves the boundary conditions for regular perturbations, and the stability conclusions obtained in the framework of GR are readily extended to the corresponding solutions of these generalized theories. In particular, we can assert that all vacuum static, spherically symmetric solutions in scalar-tensor theories, connected with the (anti-)Fisher solutions by everywhere regular conformal factors, are unstable.

However, in many cases the conformal factors bear nontrivial features, i.e., they blow up or vanish somewhere, and this can affect the boundary conditions for the perturbation equations; in any such case a separate study is necessary. Consider, for instance, the counterpart of Fisher's solution (6.54) in the Brans-Dicke (BD) scalar-tensor theory, where the Jordan-frame metric is $g^{\rm J}_{\mu\nu} = (1/\Phi) g_{\mu\nu}$; $\Phi = \exp(\phi/\sqrt{\omega + 3/2})$ is the BD scalar field, $\omega > -3/2$ is the BD coupling constant, and $|a| < 1$. It is the so-called Brans class 1 solution. In both frames, the value $u = 2k$ is a naked singularity (see more details on these solutions in [58, 85]). In the stability study, to formulate a boundary condition at this singularity, we have used the minimal requirement (6.60) for Fisher's solution , $|\delta\phi/\phi| < \infty$, providing the validity of the perturbation scheme, and we then concluded that the background solution is unstable. In the BD picture, it is more reasonable to require that the perturbed conformal factor $1/\Phi$ behave not worse than the unperturbed one, i.e., $|\delta\Phi/\Phi| < \infty$. However, since $\delta\Phi/\Phi \sim \delta\phi$, we arrive at the condition $|\delta\phi| < \infty$ which is more restrictive then (6.60) if $\phi \to \infty$, and this made us conclude in [58] that this BD solution is stable.

Another example of using the conformal mapping between Jordan and Einstein frames for stability studies can be found in [70, 71], where the instability was proved for electrically neutral and charged wormholes supported by nonminimally coupled scalar fields [22, 50, 71] (for the case of conformal coupling see Sec. 5.3). Then there is a drastic difference between the manifold structures in the two conformal frames: in the Einstein frame, without an electric charge, it is the Fisher solution with the metric $g_{\mu\nu}$ that has a singularity at $u = 2k$. In Jordan's, this singularity is removed due to the conformal factor, the solution is continued beyond this (now regular) sphere and has a flat asymptotic at the other end. Such wormholes proved to be unstable [70, 71], but their instability is of quite different nature than that described in this chapter and in [175]: it is related to a negative pole of the effective potential (6.59) at $u = 2k$ for Fisher's solution (6.54), $\epsilon = +1$ and a similar singularity in its counterpart with an electric charge. Actually, it is the same singularity as that in V_{eff} for Fisher's solution, but for the latter it occurred at the boundary of the range of the solutions whereas in continued solutions it is located somewhere in the middle.

The same reason leads to an instability [73] in the black hole solution with a massless conformally coupled scalar field [30, 41] with the metric coinciding with the extreme Reissner–Nordström metric with a scalar source — see (4.34). This solution also involves a conformally continued space-time with respect to the Einstein frame, and there is again a singularity in V_{eff} on the transition surface, but the stability study involves other boundary conditions as compared with the wormhole case since one of the boundaries $(r = m)$ is now a horizon.

Part II

Cosmology

This page intentionally left blank

Chapter 7

Stages of the Universe's evolution

This chapter describes the standard theory of our Universe's evolution. The physical essence of the phenomena under study is emphasized. For understanding the material, it is sufficient to know the fundamentals of field theory, gravitation and quantum mechanics.

7.1 The cosmological principle and the Einstein equations

When discussing many problems concerning the properties of the Universe as a whole or its part accessible to observations, one can distract from specific properties of its separate regions and the matter that fills them. It is expressed in the so-called *cosmological principle,* according to which all points of the Universe and all directions in it are equivalent, or, in other words, the space is homogeneous and isotropic. Due to the Einstein equations, the matter distribution must also be homogeneous and isotropic. The cosmological principle agrees fairly well with observations: regions of the size ≥ 100 Mpc really differ very little from each other.

The assumption of homogeneity and isotropy of space and matter distributed in it drastically simplifies the calculations and the analysis of physical processes. The corresponding equations were very briefly discussed in Chapter 2. Let us dwell upon them in more detail.

The system contains dynamic variables related to matter (density, pressure, etc.) and gravity (the metric tensor components $g_{\mu\nu}$). They obey the Einstein equations

$$G_\mu^\nu \equiv R_\mu^\nu - \frac{1}{2}\delta_\mu^\nu R = -8\pi G T_\mu^\nu - \Lambda\delta_\mu^\nu, \tag{7.1}$$

where R_μ^ν is the Ricci tensor and $R \equiv R_\alpha^\alpha$ is the scalar curvature; both of them are expressed in terms of the metric tensor and its first- and second-order derivatives with respect to the space-time coordinates. The stress-energy tensor (SET) T_μ^ν describes the properties of matter filling the space, see Chapter 2. The equations also take into account the so-called cosmological constant Λ, which is very probably observed in experiments. Eq. (7.1) can be written in another equivalent form,

$$R_\mu^\nu = -8\pi G\left(T_\mu^\nu - \frac{1}{2}\delta_\mu^\nu T\right) + \delta_\mu^\nu\Lambda, \tag{7.2}$$

where $T \equiv T_\alpha^\alpha$. Due to the contracted Bianchi identities (2.46) and the Einstein equations, the SET satisfies the equalities (2.52)

$$\nabla_\nu T_\mu^\nu = 0, \tag{7.3}$$

which have the meaning of conservation laws (see Chapter 2 and the footnote to Eq. (2.52)), and the equations of motion for matter can be derived therefrom.

Owing to the assumption of spatial homogeneity and isotropy, the squared interval can be written in the Friedmann–Robertson–Walker (FRW) form

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu = dt^2 - a^2(t)\left[\frac{dr^2}{1-kr^2} + r^2 d\Omega^2\right], \tag{7.4}$$

where $d\Omega^2 = d\theta^2 + \sin^2\theta d\varphi^2$ is the length element on a unit sphere (as in the previous chapters). Here, a particular choice of the coordinates has been made: the synchronous (true cosmological) time t and the dimensionless spherical coordinates r, θ, φ. The behaviour of the gravitational field is determined by the scale factor $a(t)$ (with the dimension of length) and the parameter k which can take three values depending on the sign of spatial curvature: $k = +1$ for spherical space, $k = 0$ for flat space, and $k = -1$ for hyperbolic space.

In the metric (7.4), the nonzero components of the Ricci and Einstein tensors are

$$R_0^0 = \frac{3\ddot{a}}{a}, \qquad R_{\underline{i}}^{\underline{i}} = \frac{1}{a^2}(a\ddot{a} + 2\dot{a}^2 + 2k),$$

$$G_0^0 = -\frac{3}{a^2}(\dot{a}^2 + k), \quad G_{\underline{i}}^{\underline{i}} = -\frac{1}{a^2}(2a\ddot{a} + \dot{a}^2 + k); \tag{7.5}$$

there is no summing over an underlined index.

Due to the symmetry properties of the tensor G_μ^ν and the Einstein equations, the tensor T_μ^ν automatically takes the form

$$T_\mu^\nu = \mathrm{diag}(\rho, -p, -p, -p), \tag{7.6}$$

where ρ and p are certain functions of the time coordinate. It is easy to verify that if the matter is a perfect fluid, then the coordinates chosen correspond to the comoving reference frame (RF) of the fluid, and the quantities ρ and p should be interpreted as its density and pressure.

Indeed, the SET of a perfect fluid is

$$T_\mu^\nu = (p+\rho)u_\mu u^\nu - p\delta_\mu^\nu, \tag{7.7}$$

where ρ and p are its density and pressure, respectively, and u^μ is the 4-velocity of its element. In the comoving RF (i.e., for an observer at rest with respect to the fluid), the 4-velocity has the form $u^\mu = (1,0,0,0)$, and the SET takes the form (7.6).

From (7.5) it is evident that there are two different Einstein equations. It is convenient to choose for consideration the equation with the left-hand side G_0^0 because it contains only a first-order derivative of the scale factor a, and the trace $R = 8\pi GT + 4\Lambda$:

$$3\frac{\dot a^2}{a^2} + \frac{3k}{a^2} = 8\pi G\rho + \Lambda, \tag{7.8}$$

$$\frac{6}{a^2}(a\ddot a + \dot a^2 + k) = 8\pi G(\rho - 3p) + 4\Lambda. \tag{7.9}$$

One more first-order equation follows from the conservation law (7.3):

$$\frac{d\rho}{da} = -\frac{1}{a}3(p+\rho),$$

or

$$\frac{d\rho}{dt} = -3H(p+\rho), \tag{7.10}$$

where the quantity $H := \dot a/a$, characterizing the expansion rate, is called the Hubble parameter.

It is easy to verify that Eq. (7.9) is a consequence of (7.8) and (7.10). Thus we have two independent equations (7.8) and (7.10) for three unknown functions $a(t)$, $\rho(t)$ and $p(t)$.

Evidently, a full solution of the equations depends on the properties of matter described by the equation of state that connects the energy density and the pressure, $p = p(\rho)$. Very frequently this relation can be chosen in the simplest linear form

$$p = w\rho, \tag{7.11}$$

where $w = \text{const}$ has different values for different kinds of matter:

$w = -1$ corresponds to a cosmological constant,
$w = 0$ to dustlike matter (or a gas of noninteracting particles),
$w = 1/3$ to disordered radiation or an ultrarelativistic gas,
$w = 1$ to matter of maximum rigidity compatible with causality, in which the velocities of sound is equal to that of light in vacuum.

Let us consider some geometric properties of FRW space-times with the metric (7.4) and an arbitrary scale factor $a(t)$.

Let one observer be located at the point $r = r_0 = 0$, and another one at a point with some $r \neq 0$. Then the last term in (7.4) makes it possible to calculate the area of the sphere $r = \text{const}$, at which the second observer is located at some time instant t: $S = 4\pi a^2(t) r^2$. Thus even if the second observer is at rest, that is, $r = \text{const}$, the physical radius of the sphere is changing proportionally to the scale factor $a(t)$:

$$R(t) \equiv a(t) r. \tag{7.12}$$

Different clusters of galaxies are examples of such "second observers". The coordinate radius of such a cluster, r, is fixed, whereas the physical radius $R(t)$, with an observer at the centre of the sphere, is growing with time. An experimental confirmation of the physical radius growth is the frequency decrease of light emitted by the second observer and received by the first one (*the cosmological redshift*).

The distances between points of space can be measured in different ways. The results of measurements (even in Minkowski space) depend on the choice of the RF and, for a time-dependent metric, even on the choice of a spatial section (i.e., on the clock synchronization method) in the framework of the same RF. We will use the notion of instantaneous physical distance, i.e., the distance at a fixed time instant in the metric (7.4). This distance is measured by a set of observers in the RF which is at rest in the metric (7.4) with clocks synchronized by the cosmological time t. At measurements along the radius r, the physical distance between neighbours separated by the coordinate difference dr is $a(t)dr/\sqrt{1-kr^2}$. The total distance

$$R_{\text{inst}} = a(t) \int_0^r \frac{dr}{\sqrt{1-kr^2}}, \tag{7.13}$$

measured by all such observers at the same time instant, is equal to the instantaneous physical radius.

If we introduce, instead of r, the coordinate χ by the equality

$$\chi = \int_0^r \frac{dr}{\sqrt{1-kr^2}} = \begin{cases} \arcsin r, & k=1, \\ r, & k=0, \\ \sinh^{-1} r, & k=-1, \end{cases} \tag{7.14}$$

we obtain in terms of χ an alternative expression for the metric (7.4):

$$ds^2 = dt^2 - a^2(t)[d\chi + r^2(\chi)d\Omega^2],$$

$$r(\chi) = \begin{cases} \sin\chi, & k=1, \\ \chi, & k=0, \\ \sinh\chi, & k=-1, \end{cases} \tag{7.15}$$

and a very simple expression for the instantaneous distance (along the radius, at fixed values of θ and φ) between the observers $r = r(\chi_1)$ and $r = r(\chi_2)$:

$$R_{\rm inst} = a(t)(\chi_2 - \chi_1). \tag{7.16}$$

Consider the motion of a light signal emitted at $t = t_0$ on an arbitrary sphere, labelled by some r, and coming at $t = t_1$ to an observer located at the origin (centre) $r = 0$. In this case the motion occurs along the radius ($d\theta = d\varphi = 0$). From the condition $ds = 0$, according to (7.4) and (7.15), we have

$$\int_{t_0}^{t_1} \frac{dt}{a(t)} = \int_0^r \frac{dr}{\sqrt{1-kr^2}} = \chi. \tag{7.17}$$

Thus the instantaneous physical distance $R_{\rm inst} = R_{\rm hor}(t_1, t_0)$ between the observer and the source (with the radial coordinate r_1 or χ_1), measured at t_1, is

$$R_{\rm hor}(t_1, t_0) = a(t_1)\chi_1 = a(t_1)\int_{t_0}^{t_1} \frac{dt}{a(t)}. \tag{7.18}$$

This quantity is called the ***particle horizon radius*** at the time t_1 for particles emitted at the time t_0. (Frequently, omitting the word "radius", it is called "the particle horizon" or even simply "the horizon" in cases where it cannot cause a confusion.)

A subtle point should be mentioned here: this distance measured at t_1 is always larger than the physical distance ***covered*** by the light signal as it travelled from the source to the observer. The latter is simply equal to $t_1 - t_0$ (or $c(t_1 - t_0)$ if one wishes to write the speed of light c explicitly), since the speed of light, measured locally at each point on the signal trajectory, is a universal constant, and the time t coincides with the physical time of the considered reference frame. The distance $R_{\rm hor}(t_1, t_0)$ is larger because the Universe has expanded while the signal travelled.[18] However, the distance $c(t_1 - t_0)$ is never used in cosmological calculations since other quantities like $R_{\rm inst}$ are much more convenient in both theoretical and observational relations.

It should be noted that the notion of a particle horizon has a few definitions at different levels of rigour. One of them is "the distance to which the light signal has moved away from its source in a certain time interval". For example, the particle horizon radius of our Universe is about 10^{28} cm for its lifetime since the beginning of expansion.

Let us also remind the reader that, apart from the notion of a particle horizon in cosmology, there are quite a number of horizons defined in the theory of gravity, which are above all used in black hole physics (see Chapters 3 and 4). Thus, the event horizon is sometimes defined as the locus of points that can be achieved by light signals, emitted from a fixed distant point, for infinite time. A typical example of such a horizon is the event horizon of the Schwarzschild BH.

Some solutions for the scale factor

Consider the time dependence of the scale factor $a(t)$ for some kinds of matter with the equation of state (7.11) in the homogeneous and isotropic space-time with the FRW metric (7.4). Recall that we can speak of spatial homogeneity and isotropy in the modern epoch only if we mean averaging of matter density at scales of the order of hundreds of megaparsecs.

The basic equations for our analysis are Eqs. (7.8) and (7.10). As will be seen from what follows, in many cases (including models of early stages of the evolution) the scale factor is much larger than unity, and the term k/a^2 in Eq. (7.8) may be neglected. Then Eq. (7.8) takes the form

$$\dot{a}^2 = \frac{8\pi G}{3}\rho a^2. \tag{7.19}$$

---

[18] The distance *covered* by the signal and the distance to which it has *moved away* in an expanding background are quite different quantities. It can be understood from the following example. Imagine that you enter an escalator moving up, ascend along it at your usual pace and have made, say, ten steps — but you are now as far away as 40 steps from the bottom at the expense of the escalator's own velocity!

Eq. (7.10) with (7.11) takes the form

$$\frac{d\rho}{da} = -3\gamma\frac{\rho}{a}, \quad \gamma := w + 1, \tag{7.20}$$

with the obvious solution

$$\rho = Ca^{-3\gamma} = Ca^{-3(w+1)}. \tag{7.21}$$

The constant C is found from the initial conditions: $C = \rho_{\rm in} a_{\rm in}^{3\gamma}$, where the subscript "in" refers to an initial instant at the evolutionary stage under consideration. Transitions between stages cannot be abrupt in reality, but this effect is often neglected.

Substituting the expression (7.21) to Eq. (7.19), it is easy to find the time dependence of the scale factor [234]:

$$a(t) = \left[a(t_{\rm in})^{3\gamma/2} + \frac{2}{3\gamma}\sqrt{\frac{8\pi G}{3}C}(t - t_{\rm in})\right]^{2/(3\gamma)}.$$

Inserting the constant C, we obtain the following expression for the scale factor with different values of w or γ:

$$a(t) = a_{\rm in}\left[1 + \frac{2}{3\gamma}\sqrt{\frac{8\pi G\rho_{\rm in}}{3}}(t - t_{\rm in})\right]^{2/(3\gamma)}. \tag{7.22}$$

The influence of the properties of matter is reflected in the parameter γ.

It is conventional to divide the Universe evolution into five stages: inflation, reheating, the radiation dominated stage, the matter dominated stage, and the modern stage of accelerated expansion, also called secondary inflation. Every stage, depending on the dominating form of matter, is approximately characterized by its parameter γ.

Since the scale factor rapidly grows with time, the first term in the brackets can frequently be neglected. As a result, we obtain the compact expression

$$a(t) \simeq a_{\rm in}\left[\frac{2}{3\gamma}H_{\rm in}\cdot(t - t_{\rm in})\right]^{2/(3\gamma)}. \tag{7.23}$$

Here $H_{\rm in} \equiv H(t_{\rm in}) = \sqrt{8\pi G\rho_{\rm in}/3}$ is the value of the Hubble parameter at the beginning of this stage.

It is easy to obtain the time dependence of the Hubble parameter:

$$H(t) \equiv \frac{\dot{a}(t)}{a(t)} = \frac{2}{3\gamma t}. \tag{7.24}$$

(We note that the Hubble parameter was called the Hubble constant for some time after the expansion of the Universe was discovered, since only times close to the present epoch were only considered, and the value of H is almost invariable during this period.)

All stages, except for the inflationary ones, are well described by the above relations. The inflationary stages, approximately described by the de Sitter metric, have a number of important features. Before discussing them, it is useful to consider the basic properties of de Sitter space.

7.2 De Sitter space

At the beginning of last century, de Sitter studied the properties of a curved space which does not contain any matter. The only feature different from Minkowski space was the presence of a nonzero cosmological constant in Eq. (7.8). The peculiar properties of this space were only of academic interest for a long time. But now it is becoming evident that it is de Sitter space that describes fairly well both the initial, inflationary stage of our Universe and the modern accelerated stage. It is therefore useful and instructive to devote some time to learning its properties.

This kind of space was briefly discussed in Chapter 2, and now we will consider it in more detail. Since there is no matter, $p = \rho = 0$, Eq. (7.8) is greatly simplified,

$$\dot{a}^2 - H^2 a^2 = -k, \quad H^2 \equiv \frac{\Lambda}{3} \tag{7.25}$$

(assuming $\Lambda > 0$) and is easily integrated:

$$k = 0: \quad a(t) = a_0\, e^{\pm Ht}, \quad a_0 = \text{const} \quad \text{(a spatially flat Universe);} \tag{7.26}$$

$$k = 1: \quad a(t) = H^{-1} \cosh[H(t - t_0)], \quad t_0 = \text{const} \quad \text{(a closed Universe);} \tag{7.27}$$

$$k = -1: \quad a(t) = H^{-1} \sinh[H(t - t_0)], \quad t_0 = \text{const} \quad \text{(a hyperbolic Universe).} \tag{7.28}$$

The choice of constants (i.e., the choice of zero point for t) such that $a_0 = H^{-1}$, $t_0 = 0$, and the plus sign in (7.26) (it corresponds to the Universe expansion) lead to a convenient general expression for an asymptotic form of all three expressions for $a(t)$ at late times:

$$a(t) = H^{-1}\, e^{Ht}, \quad t \gg H^{-1}. \tag{7.29}$$

Let us use this expression for an analysis and a deeper understanding of the general expressions obtained above. Let two observers be located at spatial points with fixed values of r: $r_1 = 0$ and $r_2 = r$, i.e., both are at rest in the RF in which the metric is written. Nevertheless, the physical distance which they measure, according to the expressions (7.12) and (7.29), grows exponentially with time.

It is also important to clearly understand the question of distances to which light signals escape in de Sitter space. Consider a flat Universe ($k = 0$) with the scale factor $a(t) = H^{-1}\,\mathrm{e}^{Ht}$ as the simplest case preserving all necessary features (by the way, all three models have the same asymptotic behaviour). For flat models, the dimensionless coordinates r and χ that we have introduced coincide. Similarly to (7.17), we obtain the expression

$$\Delta r(t, t') = \int_t^{t'} d\tau / a(\tau) = \mathrm{e}^{-Ht} - \mathrm{e}^{-Ht'} \tag{7.30}$$

for the coordinate distance covered by a light signal which has started at the time t, by the time t'. Hence the maximum value of Δr accessible to a signal emitted at the time t, is

$$\Delta r_{\mathrm{hor}} \equiv \Delta r(t,\ t' \to \infty) = \mathrm{e}^{-Ht}. \tag{7.31}$$

Thus the horizon is located at a finite coordinate distance, and even more than that, the later is the signal emitted, the smaller this distance is. In the coordinate sense, the horizon size is decreasing with time.

From the viewpoint of physical distances, the situation is different: for the time $(t' - t)$, the light signal moves away to the distance

$$R_{\mathrm{phys}}(t, t') = a(t')\Delta r(t, t') = H^{-1}[\,\mathrm{e}^{H(t'-t)} - 1].$$

With fixed initial time t, the distance grows exponentially, which is in drastic contrast with the usual situation in Minkowski space. Moreover, according to the equation

$$R_{\mathrm{hor}}(t) = R_{\mathrm{phys}}(t, t' \to \infty) = H^{-1}\,\mathrm{e}^{H(t'-t)}, \tag{7.32}$$

the horizon size tends to infinity.

The notion of a horizon is of utmost importance in cosmology because it determines the dynamic processes in the early Universe. In the inflationary period, soon after the birth of the Universe, the space-time could be approximately viewed as the de Sitter one, and the horizon size was about 10^{-27} cm. Such was the size of causally connected regions. Recalling that the Compton length of the electron is about 10^{-11} cm, it is easy to see that the very existence of the horizon inserts certain peculiarity into the physical processes. At the modern stage, dark energy is dominating, and this most probably means that we are entering a de Sitter space. In this case, the modern horizon size is $\sim 10^{28}$ cm, and larger distances are unobservable.

Is there real expansion in de Sitter space?

Let us once again consider the de Sitter metric in the form (7.4) with flat 3-space ($k = 0$):

$$ds^2 = dt^2 - H^{-2}\,\mathrm{e}^{2Ht}(dr^2 + r^2 d\Omega^2), \tag{7.33}$$

where $H = \sqrt{\Lambda/3}$. The scale factor is time-dependent, and therefore the distances between point objects at rest with respect to the cosmological RF also change.

Meanwhile, there is a coordinate transformation that brings the metric to a static form, see (3.18):

$$ds^2 = (1 - H^2 r^2)dt^2 - (1 - H^2 r^2)^{-1} dr^2 - r^2 d\Omega^2.$$

In this metric, the distances between bodies fixed in this static RF are invariable. There also exists an RF in which the distances decrease in time, e.g., this will be the case if we take the minus sign in Eq. (7.26).

Let us return to the metric (7.33) and consider test particle motion. The geodesic equation has the form

$$\frac{d^2 x^i}{ds^2} + \Gamma^i_{\mu\nu}\frac{dx^\mu}{ds}\frac{dx^\nu}{ds} = 0, \tag{7.34}$$

Since the Christoffel symbol $\Gamma^i_{tt} = 0$, one of the solutions of this equation is $x^i(t) = \text{const}$. The coordinates of points (and their differences) are time-independent. On the other hand, physical measurements ($R_{\text{phys}}(t) = \mathrm{e}^{Ht} R_{\text{phys}}(0)$) clearly indicate the growth of distances between particles at rest.

So, does the Universe really expand, while the de Sitter space is maximally symmetric and all its points-events are absolutely equivalent? This seems to mean that a transition from one time instant to another changes nothing and there is no expansion at all.

A simple explanation of this seeming paradox is that it is not the 4-dimensional space-time that expands but the 3-space of that or other RF. Moreover, we have just seen that the time lines $x^i = \text{const}$ of the RF, in which the metric has the form (7.33), are geodesics — while the distances $R_{\rm phys}$ are growing — consequently, this bundle (congruence) of geodesics is expanding. Other geodesic congruences in the same space-time can contract; it depends on the initial data specified in Eq. (7.34).

One can present one more argument confirming the reality of spatial expansion in the metric (7.33). Consider a set of test particles homogeneously distributed in the whole space and immovable in the coordinates (7.33), i.e., those with the world lines $x^i = \text{const}$. It is important that they weakly interact with each other (e.g., gravitationally). If we discover that their interaction is diminishing, then it indicates unambiguously that the distances between them grow, i.e., the Universe is really expanding.

Recall that, in the same coordinates, the horizon size (7.31) is decreasing in time by the law e^{-Ht} while the coordinate distance between the particles is fixed. Then, for any two particles there is a time instant at which the horizon size will be smaller than the distance between them, which will indicate the absence of a causal connection between the test particles. In other words, the initially nonzero interaction between the particles must tend to zero in time, hence the Universe (more precisely, the 3-space of this particular RF) is expanding.

One could present a similar reasoning in terms of physical distances, and naturally the result will be the same.

7.3 Inflation

The idea of inflation, i.e., extremely rapid expansion of the Universe at its early stage, was advanced in the early 80s of the last century. The achievements of the inflationary paradigm in explaining the observable properties of the Universe made these ideas generally accepted. A great number of inflationary scenarios are known today, and it is a hard problem to single out among them the one that was realized in the past. The number of inflationary models is continually growing. In what follows, we discuss the properties of the inflationary models based on the idea of so-called chaotic inflation.

Decades have passed since the moment when it became clear that the well-known expansion of the Universe, beginning with the "hot" (radiation-dominated) stage, leaves many questions, and it became evident that there was a nontrivial development of the Universe before the hot stage. A full list of problems can be found in [272] and in most of cosmological textbooks. The basic ones are discussed in Sec. 7.6.

The way to a solution of this set of problems was outlined in the papers [173, 191, 382].

Some time later it became clear that inflation is able to explain the basic set of observational data. The advances of the inflationary scenario were so impressive that it was included into the standard cosmological model as its main component.

The simplest and the most widespread method of describing the inflationary period is as follows. Let us assume the existence of a scalar field (inflaton) which evolves together with the gravitational field created by it. Under certain conditions, to be discussed later on, there emerges a situation resembling the de Sitter one. Thus the size of space under the horizon is growing exponentially, and it is the main feature of the inflationary period.

The Lagrangian density of a system consisting of a scalar field and gravity is written in the form

$$L = \sqrt{-g}\left\{\frac{1}{16\pi G}R + \frac{1}{2}g^{\mu\nu}\partial_\mu\varphi\partial_\nu\varphi - V(\varphi)\right\}, \tag{7.35}$$

where $g = \det(g_{\mu\nu})$ and G is the gravitational constant.

We can say in advance that the inflationary process is effective with a large energy density of the scalar field. The latter can emerge due to quantum fluctuations. Let us estimate the size of a fluctuation using the uncertainty relation

$$\Delta E \Delta t \sim 1 (= \hbar), \tag{7.36}$$

and that $\Delta E \simeq \epsilon(\Delta L)^3$. We are interested in regions where there is enough time to establish causal connections. Therefore the lifetime of such a fluctuation should not be too short. Let us suppose $\Delta t \sim \Delta L$, i.e., a light signal has time to fly through the whole region. Then the size of such a region and the energy density of the quantum fluctuation are simply related by

$$\Delta L \sim \epsilon^{-1/4}.$$

By modern views, for the beginning of inflation, the energy density must be of order of 10^{-12} in Planck units. The size of the region ΔL must be about 10^3 in Planck units, or $\sim 10^{-30}$ cm. We are coming to an understanding of the fact that quantum fluctuations of fields in the space that surrounds us permanently create regions with an increased energy density. From the viewpoint of an observer from outside, the lifetime of such a fluctuation is small. In the example under consideration it is of the order of 10^{-40} s. The spatial size of a region occupied by a fluctuation is $\sim 10^{-30}$ cm. Despite being so small, it is still larger than the Planck size, which allows for using the Einstein equations in their standard form in order to describe the processes inside such a region.

In general, a region of an increased energy density can be created artificially, for example, in heavy ion accelerators. Creation of such regions of arbitrary size, able to exist for a finite time, may become real in the future. Let such a region be formed with a size much larger than the horizon size. Let us ask ourselves: can external and internal observers make a coordinated decision on (non)expansion of this region? The answer is negative because light launched into such a region (or from it) can only cover a distance within the horizon size.

Further on, the viewpoint of an internal observer is considered.

The scalar field equation follows from (7.35):

$$\partial_\mu(\sqrt{-g}g^{\mu\nu}\partial_\nu\varphi) + \sqrt{-g}V'(\varphi) = 0. \tag{7.37}$$

The metric is assumed in the form (7.4). Due to spatial homogeneity, we also assume the homogeneity of the scalar field φ, i.e., $\varphi = \varphi(t)$. Eq. (7.37) simplifies:

$$\ddot{\varphi} + 3H\dot{\varphi} + V'(\varphi) = 0, \quad H \equiv \dot{a}/a. \tag{7.38}$$

One more equation is obtained if one takes into account that the energy density of the scalar field is $\rho = \frac{1}{2}\dot{\varphi}^2 + V(\varphi)$. Then Eq. (7.8) in the form

$$H^2 = \frac{8\pi G}{3}\left(\frac{1}{2}\dot{\varphi}^2 + V(\varphi)\right), \tag{7.39}$$

is the second equation for the set of dynamic variables $\varphi(t)$, $a(t)$. In this case we can assume that the term $\Lambda/3$ in (7.8) is either zero or is already included in the definition of the potential V.

A key point for the inflationary process is a "slow" evolution of the scalar field φ (the inflaton). In this case, the behaviour of the system (7.38), (7.39) takes place effectively in de Sitter space even if $\Lambda = 0$. Indeed, by analogy with ordinary material point mechanics, a slow motion takes place if the term responsible for friction, $3H\dot{\varphi}$, is large enough, i.e.,

$$3H|\dot{\varphi}| \gg |\ddot{\varphi}|. \tag{7.40}$$

This simplifies the set of equations even more. Indeed, using (7.40), we can write Eq. (7.38) as

$$3H\dot{\varphi} + V'(\varphi) \simeq 0. \tag{7.41}$$

Hence we arrive at the inequality

$$V'(\varphi) \sim 3H\dot{\varphi} \gg \ddot{\varphi}.$$

Multiplying the equation by $\dot{\varphi}$ and integrating, we obtain the desired inequality

$$\dot{\varphi}^2 \ll V(\varphi). \tag{7.42}$$

This inequality means that the kinetic energy is small as compared to the potential energy, and the latter therefore changes very little during the inflation $V \simeq \text{const}$, and moreover, due to Eq. (7.39), the Hubble parameter is also almost constant:

$$H \equiv \frac{\dot{a}}{a} \simeq \sqrt{\frac{8\pi G}{3} V(\varphi)}. \tag{7.43}$$

The solution to (7.43), $a(t) \propto \exp(Ht)$, means an (approximately) exponential growth of the scale factor and hence of physical distances, similar to de Sitter space. It is not surprising because an (approximately) constant potential of the inflaton φ can be interpreted as a cosmological constant.

The dynamics of a scalar field coupled to gravity is much more interesting than the dynamics of de Sitter space. In particular, the inequality (7.42) is eventually necessarily violated since the field φ, though slowly, moves towards a minimum of the potential. As soon as the potential becomes sufficiently small, the Hubble parameter H will be small as well. As a result, one can neglect the "friction" in Eq. (7.38), which means that the inflaton begins to move rapidly to the minimum of the potential, and the inflationary stage is over. Let us, however, now follow the processes at the inflationary stage. The dynamics of the inflaton φ is determined by Eq. (7.41) and that of the scale factor a by Eq. (7.43). The latter is solved in a general form:

$$a(t) = a(t_{\text{in}}) \exp\left[\int_{t_{\text{in}}}^{t} H(\varphi) dt\right]. \tag{7.44}$$

If the inflation began at $t = 0$ and has ended by $t = t_e$, then the initial spatial region has grown by a factor of $a(t_e)$. Conventionally, the Hubble parameter value $H_e \sim 10^{13}$ GeV, is used, which is in agreement with the observational data. For the overwhelming majority of estimates, it is sufficient to put $H(\varphi) \approx H_e = \text{const}$. In this case, the scale factor at the inflationary stage looks as follows:

$$a(t) \simeq a(t_{\text{in}}) \exp[H_{\text{end}}(t_{\text{end}} - t_{\text{in}})]. \tag{7.45}$$

Here H_{end} is the Hubble parameter value at the end of the inflationary period, and t_{end} is the inflation ending time.

It is also useful to introduce the notion of e-folding number, N, defined as

$$N \equiv \ln\left[\frac{a(t)}{a(t_{\rm in})}\right]. \tag{7.46}$$

In other words, if we know the number N, it means that by the present time the Universe has expanded by a factor of e^N.

If the inflation began when the scalar field had the value $\varphi_{\rm in}$, then the number N is calculated as follows:

$$N \equiv \ln\left[\frac{a(t)}{a(t_{\rm in})}\right] = \int_{t_{\rm in}}^{t} H dt = \int_{\varphi_{\rm in}}^{\varphi} H\frac{d\varphi}{\dot\varphi} = -\int_{\varphi_{\rm in}}^{\varphi} \frac{3H(\varphi)^2 d\varphi}{V'(\varphi)}. \tag{7.47}$$

Let us discuss the slow-rolling condition (7.40) in more detail. To determine more explicitly under which conditions it is possible, let us do the following. From Eq. (7.41) written in the form

$$\dot\varphi = -\frac{V'(\varphi)}{3H(\varphi)},$$

we express the second-order derivative:

$$\ddot\varphi = \frac{V'(\varphi)}{3H^2(\varphi)}H'(\varphi)\dot\varphi - \frac{V''(\varphi)}{3H(\varphi)}\dot\varphi.$$

For the slow-rolling condition to hold, both terms in the right-hand side should be small as compared to $3H\dot\varphi$. Recalling the expression (7.43) for the Hubble parameter, we obtain the following conditions:

$$\varepsilon \equiv \frac{M_{\rm Pl}^2}{16\pi}\frac{V'(\varphi)^2}{V(\varphi)^2} \ll 1; \quad \eta \equiv \frac{M_{\rm Pl}^2}{8\pi}\left|\frac{V''(\varphi)}{V(\varphi)}\right| \ll 1. \tag{7.48}$$

The stress-energy tensor (SET) of the scalar field $\varphi(t)$ in the metric (7.4), as follows from the symmetry of the problem, coincides with the SET of a perfect fluid in its comoving RF (see Sec. 5.1), $T_\mu^\nu = \mathrm{diag}(\rho, -p, -p, -p)$, and for the scalar field the energy density ρ and the pressure p are expressed as

$$\rho = \frac{1}{2}\dot\varphi^2 + V(\varphi) \tag{7.49}$$

$$p = \frac{1}{2}\dot\varphi^2 - V(\varphi). \tag{7.50}$$

In the inflationary period, $\frac{1}{2}\dot\varphi^2 \ll V(\varphi)$, and this means an approximately de Sitter space with the corresponding behaviour of the scale factor.

7.4 Post-inflationary stages

7.4.1 Post-inflationary reheating of the Universe

The stage right after inflation is probably the most difficult for analysis. Even its name (*reheating*) is not quite correct since there has not been any primary heating. It is this stage during which active high-energy particle production and their thermalization occurs, which, in the language of statistical physics, means heating of the plasma just born.

The inflationary period terminates when the "friction" becomes small, i.e., at $H_{\text{end}} \lesssim m_{\text{infl}}$ (m_{infl} is the inflaton mass). The inflaton field begins to rapidly oscillate near the minimum of the potential according to Eq. (7.38). It is this motion, alternating with time, that creates the particles. The oscillation energy gradually passes over to the energy of particles. Using the expressions for the pressure, $p = \frac{1}{2}\dot{\varphi}^2 - V(\varphi)$, and for the energy density, $\rho = \frac{1}{2}\dot{\varphi}^2 + V(\varphi)$, it is easy to transform this equation to the familiar form

$$\frac{d\rho}{dt} = -3H(t)(p+\rho). \tag{7.51}$$

The additional condition (the equation of state) $\rho + p = \gamma\rho$, holds only approximately, after averaging over the oscillation period. Meanwhile, the parameter γ depends on the model:

$$\gamma \simeq \frac{\int\limits_0^T (\rho+p)dt}{\int\limits_0^T \rho dt} = \frac{\int\limits_0^T \dot{\varphi}^2 dt}{\int\limits_0^T \left[\frac{1}{2}\dot{\varphi}^2 + V(\varphi)\right] dt} = \frac{\int\limits_0^{\varphi_{\max}} \dot{\varphi} d\varphi}{\int\limits_0^{\varphi_{\max}} d\varphi \left[\frac{1}{2}\dot{\varphi}^2 + V(\varphi)\right]/\dot{\varphi}}. \tag{7.52}$$

This expression can be further simplified if one assumes that, during a single oscillation period, the Universe expands very little, which is quite realistic. Then, according to the energy conservation law, we have $\dot{\varphi} = \sqrt{2[V(\varphi_{\max}) - V(\varphi)]}$. Near its minimum, the potential is approximated by the parabola $V(\varphi) \cong \lambda\varphi^2$, Eq. (7.52) is strongly simplified, and the quantity γ at the reheating stage turns out to be equal to unity,

$$\gamma \cong 1. \tag{7.53}$$

Let us determine the time dependence of the scale factor. Its initial value at the reheating stage coincides with that at the end of inflation, $a(t_{\text{in}}) = a_{\text{infl}}(t_{\text{end}}) = H(\varphi_U)^{-1}\,\mathrm{e}^{N_U}$. It is assumed that the visible part of the Universe has formed at N_U e-foldings before the end of inflation. The initial size of the Universe was equal to the inverse Hubble parameter, $H(\varphi_U)^{-1}$, while the inflaton value was φ_U. The energy density at the beginning of reheating was the same as at the end of inflation, $\rho(t_{\text{in}}) = \rho_{\text{infl}}(t_{\text{end}}) = V(\varphi_{\text{end}})$. Then Eq. (7.23) is written as

$$a_{\text{reh}}(t) = \left(\frac{2}{3}H_{\text{end}}\right)^{2/3} H(\varphi_U)^{-1}\, \mathrm{e}^{N_U}\, t^{2/3}. \tag{7.54}$$

Recall that it is usually assumed that $H_{\text{end}} \simeq H(\varphi_U) \simeq 10^{13}$ GeV.

During the rapid oscillations, the inflaton field produces particles, giving them a part of its energy. Suppose that the interaction energy density of the inflaton (φ) with fermions ψ and other scalar particles χ has the usual form,

$$L_{\text{int}} = h\varphi\bar{\psi}\psi + g\varphi\chi^2,$$

where g and h are coupling constants. Then the rate of change of inflaton energy density is

$$\dot{\rho} = \frac{d}{dt}\left[\frac{1}{2}\dot{\varphi}^2 + V(\varphi) + \Delta L\right] = \dot{\varphi}\ddot{\varphi} + \dot{\varphi}V'(\varphi) + \Delta\dot{L},$$

where ΔL is a term responsible for diminishing the inflaton energy density at its decay into other, secondary particles.

Suppose that the back reaction of the radiation on the inflaton dynamics is small [141]. Then, in the first approximation, according to the equations of motion, we obtain

$$\ddot{\varphi} \simeq -3H\dot{\varphi} - V'.$$

Substituting into the previous expression, we find the changing rate of the inflaton energy density

$$\dot{\rho} \simeq -3H(\rho + p) + \Delta\dot{L}.$$

Now the meaning of the last term is clear: it is the changing rate of inflaton energy density due to decay into other particles,

$$\Delta\dot{L} = -\Delta\dot{\rho},$$

where $\Delta\dot{\rho}$ is the secondary particles' energy density changing rate due to the inflaton decay. Expanding it in a Taylor series by the derivative of the field φ, we have

$$\Delta\dot{\rho}(\dot{\varphi}) = \Delta\dot{\rho}(0) + A\dot{\varphi} + \Gamma\dot{\varphi}^2 + \ldots \simeq \Gamma\dot{\varphi}^2.$$

The first term of the expansion is zero because a static field does not radiate, and the second one is zero because at oscillations of the inflaton the mean value of its derivative is (approximately) zero. Let us also take into account that the pressure and energy density of the oscillating inflaton field satisfy the usual relation $p = w\rho$. Then,

$$\dot{\rho} \simeq -(3H + \Gamma)\dot{\varphi}^2 = -(3H + \Gamma)(p + \rho) \simeq -(3H + \Gamma)(w + 1)\rho.$$

Recalling that $H = \dot{a}/a$, we obtain the time dependence of the inflaton energy density as

$$\rho = \rho_{\text{end}} \left(\frac{a(t)}{a(t_{\text{end}})} \right)^{-3(w+1)} \exp\{-\Gamma(w + 1)(t - t_{\text{end}})\}, \tag{7.55}$$

where the subscript "end" refers to the inflation ending time. This expression is quite understandable physically. The second factor means that the inflaton energy density is decreasing due to expansion of space, and the third one due to its decay into particles.

What is now happening to the energy density of the created particles of matter? The changing rate of the energy density of relativistic particles ρ_m is written as follows:

$$\begin{aligned} \dot{\rho}_m &= -3H(p_m + \rho_m) + \Gamma\dot{\varphi}^2 \\ &= -3H(w_m + 1)\rho_m + \Gamma(w + 1)\rho \\ &= 4H\rho_m + \Gamma(w + 1)\rho. \end{aligned} \tag{7.56}$$

Here it has been taken into account that $w_m = 1/3$ for relativistic particles. It is evident that two terms are competing: the density decrease due to expansion of space (the first term) and its increase due to inflaton decay.

The relativistic particles rapidly thermalize. When the oscillations begin, the energy inflow into the particle plasma is predominant, and the temperature is growing. As time passes, the oscillation intensity decreases, and the space expansion effect that cools the plasma becomes more important. Let us estimate the temperature of the plasma at which its density is the highest. Assuming $\rho_m \sim \rho$, we obtain from the previous equation $H \sim \Gamma$. Since the Hubble parameter and the cosmological time are related by $t \sim 1/H$ (see (7.24)), while the temperature is related to time as $T \sim \sqrt{M_{\text{Pl}}/t}$ (see Sec. 7.4.2 and Eq. (7.63) for more detailed consideration), we arrive at the following estimate:

$$T_{\text{reh}} \sim \sqrt{M_{\text{Pl}}/t} \sim \sqrt{M_{\text{Pl}}H} \sim \sqrt{M_{\text{Pl}}\Gamma}.$$

A large uncertainty in the calculation of the quantity Γ prevents us from determining the heating temperature unambiguously. It is usually supposed that it varies in a wide range between 10^4 and 10^{12} GeV.

7.4.2 The radiation-dominated stage

Up to now, we succeeded in expressing the energy density in terms of fields, and this allowed us to solve Eq. (7.41) and find the time dependence of the scale factor. But, by the end of the reheating stage, the inflaton oscillations have decayed, having created high-energy particles. The existence of the particle plasma indicate that it is desirable to introduce the notion of temperature, to use the standard tools of statistical physics and thermodynamics. The notion of temperature is only rigorous if the system is in equilibrium. It our case it is evidently not true because the Universe is expanding.

Nevertheless, the relaxation processes occur so rapidly that at each time instant the state of the Universe is close to equilibrium. To verify that, let us assume that it is the case and check if we come to a contradiction. We introduce the plasma temperature T at the time t. It will be shown below that the cosmological time and the temperature are related by $t \sim M_{\rm Pl}/T^2$ (see Eq. (7.63)). An equilibrium is established due to particle collisions. The characteristic time of a collision between an electron and a photon, $t_{\gamma e}$, is $t_{\gamma e} \sim 1/(n\sigma v)$, where $n \sim T^3$ is the electron density, $\sigma \sim \alpha^2/T^2$ is the Compton cross-section, and $v \simeq 1$ is the electron velocity. We can use the notion of temperature only if $t_{\gamma e} \ll t$. Substituting the estimates presented in this paragraph, we obtain the applicability condition of the notion of temperature in the expanding Universe in the form $T \ll \alpha^2 M_{\rm Pl} \sim 10^{17}$ GeV. It is known that the temperature at the end of inflation can hardly exceed 10^9 GeV. Thus an equilibrium in the medium is established very rapidly, at least due to Compton scattering, which justifies introducing the notion of temperature.

To find an explicit form of the scale factor $a_{\rm RD}(t)$ at the radiation-dominated stage, we use the general relation (7.22) with $\gamma = 4/3$. This value follows from the pressure-density relation $p = \rho/3$ for a medium consisting of relativistic particles and Eq. (7.20).

The initial conditions for this stage are $t_{\rm in} = t_{\rm reh}$, $\rho_{\rm in} = \rho(t_{\rm reh})$ and $a_{\rm in} = a(t_{\rm reh})$, where the subscript 'reh' refers to the end of the previous stage, reheating . The final expression for the scale factor is

$$a(t) \equiv a_{\rm RD}(t) = a_{\rm reh}(t_{\rm reh}) \left[\frac{1}{2} H(t_{\rm reh})\right]^{1/2} (t - t_{\rm reh})^{1/2}. \qquad (7.57)$$

Let us look how the temperature evolves with time. First of all, we recall that the energy density of the ultrarelativistic particles is related to the temperature T by

$$\rho = \frac{\pi^2}{30} g_* T^4, \tag{7.58}$$

where g_* is the number of types of particles, taking into account their statistical weight.

Let us further establish the important relationship between the temperature and the scale factor. On the one hand, at this stage $\gamma = 4/3$, and according to (7.21) we have $\rho \sim a^{-4}$. On the other hand, $\rho \sim T^4$. Therefore,

$$a(t) = \text{const}/T(t). \tag{7.59}$$

The temperature decreases inversely proportionally to the scale factor.

Let us show that the entropy after the reheating stage (when the particles were created and hence the entropy grew) remains constant. Indeed, using the known relationship between the entropy density and the temperature

$$s = \frac{2\pi^2}{45} g_* T^3, \tag{7.60}$$

it is easy to obtain

$$S \sim sa^3(t) = \text{const}. \tag{7.61}$$

Another useful relation, following from (7.59), is also important for making estimates in cosmological models:

$$T(t) = \frac{a(t_{\text{reh}})}{a(t)} T_{\text{reh}}. \tag{7.62}$$

Lastly, combining Eqs. (7.57), (7.58) and (7.59), it is easy to obtain the explicit time dependence for the temperature of the medium filling the Universe:

$$T = \left(\frac{45}{32\pi^2 g_*}\right)^{1/4} \sqrt{\frac{M_{\text{Pl}}}{t}}. \tag{7.63}$$

Let us stress that this relation is valid in the radiation-dominated stage only.

7.4.3 The matter-dominated stage

As the Universe expands, the medium temperature falls, massive particles become nonrelativistic, and the photon wavelengths grow. There comes a moment when the particles' rest energy becomes equal to their kinetic energy. Beginning with this moment, there begins the matter-dominated stage. From that time on, the pressure p becomes so small that it can be put equal to zero, and accordingly the parameter γ is equal to 1 in Eq. (7.20). By full analogy with the previous stages, the scale factor $a_{\text{MD}}(t)$ can be found from Eq. (7.22) with $\gamma = 1$. The initial conditions for this stage are

$$t_{\text{in}} = t_{\text{RD}}, \quad \rho_{\text{in}} = \rho(t_{\text{RD}}) = \frac{\pi^2}{30} g_* T_{\text{RD}}^4, \quad a_{\text{in}} = a_{\text{RD}}(t_{\text{RD}}),$$

where t_{RD} corresponds to the end of the radiation-dominated stage.

It is of interest that Eq. (7.59) remains valid at this stage as well, with one reservation: now T is the photon temperature. The latter are called relic photons.

Before beginning to prove (7.59), let us discuss one more period of utmost importance in the life of the Universe, the so-called recombination period. The point is that at high temperatures electrons and protons cannot form hydrogen atoms because of their small binding energy. The Universe is filled with a plasma of charged particles which actively interact with photons. It means, by the way, that the emitted photons performed something like Brownian motion. The Universe was not transparent to electromagnetic radiation. But at a certain moment, when the temperature became low enough, hydrogen atoms began to form. Being neutral, they weakly interacted with photons allowing them to propagate to large distances. The time interval during which the atoms recombined and the medium was becoming transparent is called the recombination period. All the existing data say that the two periods, that of recombination and that of equality between the energy densities of matter and radiation, roughly coincided in time. In what follows, we will neglect their difference, which is small from the cosmological viewpoint.

Thus we suppose that the photons do not interact with matter after reaching the temperature T_{rec}. Before this time, the photon distribution by energies was a Bose-Einstein distribution, which, by the recombination time, had the form

$$dN(t_{\text{rec}}) = V_{\text{rec}} \frac{E_{\text{rec}}^2}{\pi^2} \frac{dE_{\text{rec}}}{\exp(E_{\text{rec}}/T_{\text{rec}}) - 1}. \tag{7.64}$$

Here, dN is the number of photons with energies between E_{rec} and $E_{\text{rec}} + dE_{\text{rec}}$ inside the volume V_{rec}. Since after recombination the photons' interaction with the ambient medium is regarded to be small, the photon energy decreases only due to the common expansion of space,

$$E(t) = p(t) = \frac{2\pi}{\lambda(t)} = \frac{2\pi}{\dfrac{a(t)}{a(t_{\rm rec})}\lambda(t_{\rm rec})} = \frac{a(t_{\rm rec})}{a(t)} E_{\rm rec}. \tag{7.65}$$

The number of photons is a conserved quantity, so that

$$dN(t) = dN(t_{\rm rec}),$$

while the volume V containing these particles grows by the law

$$V(t) = \left(\frac{a(t)}{a(t_{\rm rec})}\right)^3 V_{\rm rec}.$$

As a result, we obtain the photon distribution at arbitrary time t:

$$dN(t) = V(t)\frac{E(t)^2}{\pi^2}\frac{dE}{\exp\left[E(t)\dfrac{a(t)}{a(t_{\rm rec})}\Big/ T_{\rm rec}\right] - 1}. \tag{7.66}$$

The distribution preserves its Bose-Einstein form, and

$$T(t) = \frac{a(t_{\rm rec})}{a(t)} T_{\rm rec}. \tag{7.67}$$

Our statement has been proved: the law (7.59) really holds at both the radiation-dominated stage and the matter-dominated stage. The unknown constant in (7.59) can be determined from normalization for the present time: const $= a_0 T_0$, or for the recombination time, as has been done before.

7.4.4 The modern stage of accelerated expansion (secondary inflation)

More and more confidently, the observational data show that the density of dark energy, a substance uniformly distributed in space, contributes to the modern total energy density at about 70 per cent. The most natural assumption is that this quantity is time-independent and is related to the cosmological constant Λ. This means in turn that in the present epoch the Universe is approximately described by a de Sitter model but with the energy density much smaller than in the inflationary period. More than that, with further expansion of space, the matter density decreases while the energy density related to the cosmological constant remains constant. Consequently, the Hubble parameter also tends to a constant value,[19]

$$H_\Lambda \to \sqrt{\frac{8\pi}{3}\frac{\rho_\Lambda}{M_{\rm Pl}^2}}, \tag{7.68}$$

and the scale factor to the function

$$a(t) = H_\Lambda^{-1}\,{\rm e}^{H_\Lambda(t-t_0)}. \tag{7.69}$$

The Hubble parameter characterizes the rate of expansion, and to describe the change of this rate, i.e., the acceleration, one more dimensionless parameter is used, the deceleration parameter $q(t) := -a\ddot{a}/\dot{a}^2$. The parameter q was introduced when it was believed that the Universe is expanding with deceleration. Accelerated expansion corresponds to values $q < 0$.

The modern observational estimates of these parameters are different as given by different authors but are approximately in the following limits:

$$H_0 \approx 0.71 \pm 0.04\,\frac{\rm km}{{\rm c}\cdot{\rm Mpc}}, \quad q_0 \approx -1 \pm 0.4, \tag{7.70}$$

where the subscript "0" corresponds to the modern epoch. In the case (7.69) we obtain $q(t) = q_0 = -1$.

The inverse Hubble parameter characterizes the size of a causally connected region. An estimate on the basis of the observational data gives

$$H^{-1} \approx 10^{28}\,{\rm cm} \approx 10^{62} M_{\rm Pl}.$$

It is this quantity that gives the size of the visible part of the Universe. If the Hubble parameter is really constant, then information on more remote parts of the Universe will never be available to us.

When using relations like (7.22) [such as Eqs. (7.77), (7.78), (7.57), (7.80) given below] for obtaining numerical estimates, there emerges a question connected with the particular value of the scale factor $a_{\rm in}$ at the beginning of each stage. Besides, the value of the coordinate distance r is not evident. Therefore a problem appears when calculating, e.g., the

[19]The cosmological constant corresponds to a density to pressure ratio w equal to negative unity. The modern cosmological observations admit a certain interval of w values including -1. Let us present, as examples, two recent estimates: $w = -1.10 \pm 0.14(1\sigma)$ [245] (according to the 7-year WMAP data), and $w = -1.069^{+0.091}_{-0.092}$ [389] (mainly from data on type Ia supernovae from the SNLS3 sample). Noteworthy, the central values of w in these estimates belong to the phantom range, $w < -1$.

scale factor $a(t_0)$ at a given time instant t_0 and consequently the physical distance between two objects, equal to

$$R_0 = a(t_0)r. \tag{7.71}$$

It is this quantity that is measured by observers. The problem can be circumvented if one considers the expression (7.71) as a way to get rid of the comoving coordinate r by normalizing the scale factor to its present value $a_0 = a(t_0)$. Then the distance between two point objects at arbitrary time is expressed as

$$R(t) = \frac{a(t)}{a_0} R_0.$$

For example, the distance during the matter-dominated stage is

$$R(t) = \left(\frac{t - t_{\rm RD}}{t_0 - t_{\rm RD}}\right)^{2/3} R_0 \simeq \left(\frac{t}{t_0}\right)^{2/3} R_0.$$

The lifetime of the Universe t_0 and its size a_0 are already well defined. The dimensionless scale factor is also frequently introduced:

$$\mathrm{a}(t) \equiv \frac{a(t)}{a_0}, \quad 0 < \mathrm{a}(t) \leq 1.$$

7.4.5 Future of the Universe: Is a Big Rip expected?

The description of dark energy with the aid of the cosmological constant in the Einstein equations well agrees with observations in the modern epoch. However, the same observations leave quite a wide range of admissible values of the parameter w if one describes dark energy as a perfect fluid with the equation of state $p = w\rho$. For such a description, in general, the parameter w is time-dependent, while for $w = \mathrm{const}$ the conservation law leads to Eq. (7.21), $\rho \sim a^{-3(w+1)}$.

It is easy to verify that an accelerated expansion, $q < 0$, requires a negative pressure, $w < -1/3$. Then, for a large scale factor a, the term with k in Eq. (7.82) is small relative to the other terms, so that ***the late-time behaviour of the scale factor does not depend on the spatial curvature***, and the dependence $a(t)$ can be described as follows:

(a) For $-1/3 > w > -1$, so that $\rho > |p|$ (the dominant energy condition holds), the so-called power-law inflation occurs:

$$a \sim t^{2/[3(w+1)]}, \quad q = -1 + \frac{3}{2}(w+1) > -1; \tag{7.72}$$

(b) at $w = -1$, which corresponds to the cosmological constant $\Lambda > 0$, $\rho = \text{const} > 0$, exponential inflation is obtained:

$$a \sim e^{Ht}, \quad H = \text{const}, \quad q = -1; \tag{7.73}$$

(c) for $w < -1$ with the so-called phantom matter, hyperinflation takes place, ending with a singularity due to growth of the scale factor:

$$a \sim (t_* - t)^{-2/[3|w+1|]}, \quad q = -1 - \frac{3}{2}|w+1| < -1, \tag{7.74}$$

where t_* is the time of a singularity called the Big Rip.

As is clear from the estimates (7.70), from the observational viewpoint, all three variants are admissible. However, since the acceleration not only exists but increases, there is some preference for (b), and many papers discuss cosmological models where the dark energy "overcomes the phantom barrier", i.e., at some instant ($3 - 4 \cdot 10^9$ years ago) passes over from $w > -1$ to $w < -1$.

In case (c) matter behaves quite exotically: unlike usual matter, its density does not decrease as the volume grows, but even grows, and at the singular time both a and ρ blow up simultaneously.

The catastrophic growth of $a(t)$ makes all distances grow, beginning with extragalactic ones, then also interstellar, interplanetary, and finally intra-atomic ones, i.e., right before the singularity all kinds of matter should decay and even all composite particles.

Is such a sad fate of our Universe inevitable if a modern value $w < -1$ is after all confirmed? Fortunately, the answer is negative.

Indeed, if $w = p/\rho$ is time-dependent, one can assume that the dark energy is represented by a scalar field ϕ with the Lagrangian

$$L_s = \frac{1}{2}\varepsilon g^{\mu\nu}\phi_{,\mu}\phi_{,\nu} - V(\phi), \tag{7.75}$$

where $\varepsilon = \pm 1$ and $V(\phi)$ is the self-interaction potential of the field ϕ. These are the same normal and phantom scalar fields that we met in the previous chapters. At $\phi = \phi(t)$,

$$\rho = \frac{1}{2}\varepsilon\dot{\phi}^2 + V, \quad p = \frac{1}{2}\varepsilon\dot{\phi}^2 - V, \quad w = \frac{p}{\rho} = -1 + \frac{2\varepsilon\dot{\phi}^2}{2V + \varepsilon\dot{\phi}^2}. \tag{7.76}$$

Thus a normal scalar field ($\varepsilon = +1$) with a positive potential V gives $w > -1$ while a phantom scalar field ($\varepsilon = -1$) with $V > 0$ leads to $w < -1$. However, if at large t the scalar field tends to an extremum of the potential $V_{\text{ext}} > 0$ sufficiently rapidly, then $\dot{\phi} \to 0$ and $w \to -1$ as $t \to \infty$; V_{ext}

behaves as an effective cosmological constant, and accordingly we obtain a de Sitter asymptotic (7.73) for the Universe evolution.

It is well known that a normal scalar field tends to a minimum of its potential ("rolls down" along the curve $V(\phi)$), whereas a phantom one, on the contrary, tends to a maximum of V ("climbs up" along the curve $V(\phi)$). If $V(\phi)$ has a maximum, then a phantom scalar field will tend to it in the course of the cosmological evolution. For this case, an exact result is known [160]: *if the potential $V(\phi)$ is bounded above, then its maximum represents a global attractor for cosmological solutions to the gravitational and scalar field equations.*

7.5 The scale factor in the general case

It is useful to put together the basic expressions for the scale factor at different stages. Since each stage is much longer than the previous ones, i.e., $t \gg t_{\rm in}$, we use the more compact but approximate equation (7.23).

The inflationary stage:

$$a(t) \equiv a_I(t) = H_U^{-1} \exp\left(\int H\,dt\right) \approx H_e^{-1}\,{\rm e}^{N_U t}. \tag{7.77}$$

The numerical values $N_U \approx 60,\ H_e \approx 10^{13}$ GeV agree with observations for the majority of models.

The reheating stage:

$$a(t) \equiv a_{\rm reh}(t) = a_I(t_e)\left(\frac{2}{3}H_e\right)^{2/3}(t-t_e)^{2/3}. \tag{7.78}$$

The radiation-dominated stage:

$$a(t) \equiv a_{\rm RD}(t) = a_{\rm reh}(t_{\rm reh})\left[\frac{1}{2}H(t_{\rm reh})\right]^{1/2}(t-t_{\rm reh})^{1/2}. \tag{7.79}$$

The matter-dominated stage:

$$a(t) \equiv a_{\rm MD}(t) = a_{\rm RD}(t_{\rm RD})\left[\frac{2}{3}H(t_{\rm RD})\right]^{2/3}(t-t_{\rm RD})^{2/3}. \tag{7.80}$$

The modern accelerated expansion (secondary inflation), assuming that the dark energy is described by the cosmological constant Λ in the Einstein equations:

$$a(t) = H_\Lambda^{-1}\,{\rm e}^{H_\Lambda(t-t_0)}, \quad H_\Lambda = 3/\Lambda. \tag{7.81}$$

All these expressions are approximate because they take into account only one predominant source of the gravitational field at each stage.

The real content of the Universe is a mixture of radiation, matter and dark energy which is rather probably described by the cosmological constant. Let us obtain a general expression for the scale factor, using Eq. (7.8):

$$\frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = \frac{8\pi G}{3}\rho + \frac{\Lambda}{3}. \tag{7.82}$$

The contribution from the dark energy density ρ_Λ is written separately as the Λ-term, $\Lambda/3 = (8\pi G/3)\rho_\Lambda$. The remaining part of the energy density consists of the matter density ρ_m and the radiation density ρ_r,

$$\rho = \rho_m + \rho_r. \tag{7.83}$$

Their dependence on the scale factor has already been discussed,

$$\rho_m(t) = \rho_m(t_0)/a^3(t), \quad \rho_r(t) = \rho_r(t_0)/a^4(t).$$

Substituting these quantities into (7.82), we obtain the equation for the scale factor

$$\frac{\dot{a}^2}{a^2} = \frac{8\pi G}{3}\left[\frac{\rho_m(t_0)}{a^3(t)} + \frac{\rho_r(t_0)}{a^4(t)} + \rho_\Lambda\right] - \frac{k}{a^2}. \tag{7.84}$$

Let us choose the additional condition in the form $a(t_0) = 1$. Numerical solution of this equation does not represent any problem. Let us transform it, introducing some definitions often used in practice.

First of all, we define the critical density at the modern epoch:

$$\rho_c \equiv \frac{3H_0^2}{8\pi G}. \tag{7.85}$$

where H_0 is the modern value of the Hubble parameter. At such a value of the energy density the Universe should be flat ($k = 0$) provided $\Lambda = 0$. In general, the mean density of the Universe $\rho \neq \rho_c$, therefore it is convenient to introduce the dimensionless density parameter

$$\Omega_{\text{tot}} = \rho/\rho_c. \tag{7.86}$$

Let us also introduce the partial density parameters Ω for the present time. Thus, for matter, radiation and the cosmological constant we have

$$\Omega_m = \frac{\rho_m(t_0)}{\rho_c}, \quad \Omega_r = \frac{\rho_r(t_0)}{\rho_c}, \quad \Omega_\Lambda = \frac{\rho_\Lambda(t_0)}{\rho_c}.$$

An expression for Ω_k, the density parameter related to the spatial curvature is constructed in a different way:

$$\Omega_k = -\frac{k}{{H_0}^2}.$$

Now Eq. (7.84) looks as follows:

$$\frac{\dot{a}^2(t)}{a^2(t)} \equiv H^2(t) = H_0^2[\Omega_k a^{-2}(t) + \Omega_m a^{-3}(t) + \Omega_r a^{-4}(t) + \Omega_\Lambda], \quad (7.87)$$

where $H_0 = H(t_0)$. Choosing $t = t_0$ in equation (7.87), we obtain the sum rule

$$\Omega_r + \Omega_m + \Omega_\Lambda + \Omega_k = 1.$$

Lastly, let us present the time dependence of the horizon size (in the present case, it means the distance to which a light signal can get for the time t):

$$R_{\text{hor}}(t) = a(t) \int_0^t \frac{dt'}{a(t')}.$$

At the inflationary stage, $a(t) = \text{const}\, e^{Ht}$, where H is the Hubble parameter expressed in terms of energy density and assuming a constant value with good accuracy. The horizon size is $R_{\text{hor}} = H^{-1}(1 - e^{-Ht})$, and at large times it tends to the inverse Hubble parameter, i.e., also a constant.

At other stages, we have the approximation $a(t) = \text{const} \cdot t^\beta$, where the constant $\beta < 1$. Then the Hubble parameter has the form $H(t) = \beta t^{-1}$, and the horizon size is $R_{\text{hor}} = t/(1-\beta)$. Thus $H^{-1} = R_{\text{hor}}(1-\beta)/\beta$, i.e., the inverse Hubble parameter coincides with the horizon size by order of magnitude.

7.6 Why do we need an inflationary period?

In scenarios of the Universe's birth studied before the inflationary idea was put forward, the first stage was that of a "hot" radiation-dominated Universe. All of them face some unsolvable problems. Let us enumerate the most significant ones.

1. The initial singularity. A study of the classical equations of GR applied to the Universe as a whole leads to the conclusion that its initial state was characterized by an infinite density. The inflationary idea does not completely solve this problem though indicates a way of its solution. Very probably, our Universe has originated from quantum fluctuations where the classical equations, and, the more so, the conclusions that rest on them, were unapplicable.

2. Our space is described with high accuracy by Euclidean geometry. Its causes are not evident.
3. Our Universe contains about 10^{22} stars. How could an object of such mass and volume emerge? Why does it exist for such a long time?
4. The cosmic microwave background (CMB) is highly isotropic. Why are its energy density deflections so small?
5. How did the energy density inhomogeneities that have led to galaxy formation appear?
6. Why does matter dominate in our Universe while the laws of Nature are almost symmetric with respect to changing particles to antiparticles?
7. At the early stage, at large energy densities, topologically stable states like magnetic monopoles should form. Their number and mass are large enough to be noticeable [436]. But such objects have not yet been discovered.

Let us discuss the main problems. The first of them is the Universe flatness problem.

7.6.1 The flatness problem

Let us recall the notion of the critical density defined as the density of matter such that the curvature parameter is $k = 0$ at $\Lambda = 0$. From Eq. (7.8) it directly follows

$$\rho_{\text{crit}}(t) \equiv \frac{3}{8\pi} M_{\text{Pl}}^2 H^2(t). \tag{7.88}$$

(Keep in mind that $G = 1/M_{\text{Pl}}^2$ and $\dot{a}/a = H$.) Since the constant Λ is small at least at the radiation-dominated stage, let us neglect it in Eq. (7.8), which results in

$$\left| \frac{\rho_{\text{crit}}(t) - \rho(t)}{\rho_{\text{crit}}(t)} \right| = \dot{a}^{-2}(t), \tag{7.89}$$

for $k = \pm 1$. The modern value of the energy density is very close to the critical value, [379], so that the inequality

$$\left| \frac{\rho_{\text{crit}}(t_0) - \rho(t_0)}{\rho_{\text{crit}}(t_0)} \right| = \dot{a}^{-2}(t_0) < 1, \tag{7.90}$$

is established confidently. From Eqs. (7.89) and (7.90) we obtain

$$\left| \frac{\rho_{\text{crit}}(t) - \rho(t)}{\rho_{\text{crit}}(t)} \right| = \left| \frac{\rho_{\text{crit}}(t_0) - \rho(t_0)}{\rho_{\text{crit}}(t_0)} \right| \frac{\dot{a}^2(t_0)}{\dot{a}^2(t)} < \frac{\dot{a}^2(t_0)}{\dot{a}^2(t)}.$$

In the absence of inflation, there are only two stages, the radiation- and matter dominated ones. The time dependence of the scale factor at both stages can be presented in the form

$$a(t) = \text{const} \cdot t^n, \quad n = 1/2 \quad \text{or} \quad 2/3 < 1. \tag{7.91}$$

Consequently,

$$\left| \frac{\rho_{\text{crit}}(t) - \rho(t)}{\rho_{\text{crit}}(t)} \right| < \frac{\dot{a}^2(t_0)}{\dot{a}^2(t)} \sim \left(\frac{t}{t_0} \right)^{2(1-n)} \ll 1. \tag{7.92}$$

Then, if the dependence (7.91) is correct, at birth of the Universe, i.e., at $t \ll t_0$, the matter density must be equal to the critical one with inexplicable accuracy. The only way out from this situation is a violation of the law (7.91), at least at the initial stage. It is this way out that is provided by inflation.

Let the Universe pass an inflationary stage with the scale factor (7.44) $a(t) = H^{-1}\cosh(Ht)$. The Hubble parameter at inflation is almost constant, $H \simeq \text{const}$. Then the time dependence of the ratio (7.89) is

$$\left| \frac{\rho_{\text{crit}}(t) - \rho(t)}{\rho_{\text{crit}}(t)} \right| = \sinh^{-2}(Ht).$$

This function exponentially decreases at the inflationary stage, and this explains its smallness in the post-inflationary period, see (7.92).

7.6.2 The initial size of the Universe

If now the Universe has the size $L_U \sim 10^{28}$ cm, at any t its size is

$$L(t) = \frac{a(t)}{a(t_0)} L_U = \frac{T_0}{T} L_U. \tag{7.93}$$

Let us estimate the smallest size of the Universe. From Eq. (7.93) it follows that it took place at the greatest temperature. The classical description is valid at energies, hence temperatures smaller than the Planckian energy, $T \leq M_{\text{Pl}} \sim 10^{32}$ K. Substituting this value into (7.93) and taking into account the modern CMB temperature $T_0 \sim 2.7$ K, we obtain the minimum size of the Universe at birth:

$$L_{\min} \sim 10^{-4}\,\text{cm}.$$

It remains an enigma why was the Universe born with such a large size as compared to the Planckian one.

The inflationary mechanism just allows the Universe to rapidly expand to the above size.

Moreover, if the Universe was born with the size $\sim 10^{-4}$ cm at the Planckian energy scale, then, at the same moment, the size of a causally connected region is about 10^{-33} cm. It means that the Universe consisted of $\sim 10^{87}$ causally disconnected regions. It is absolutely unclear why the energy density is the same in all these regions up to about 10^{-5}. It is the essence of the ***horizon problem***.

The assumption of the existence of an inflationary period solves this and other problems, such as the ***monopole problem*** [436].

7.6.3 Causal connections at inflation and after it

The inflationary stage has a feature of great interest, distinguishing it from all other stages. Since an understanding of this feature is necessary for understanding the majority of subsequent processes, let us dwell upon it in more detail.

Consider two spatial points at rest in the cosmological RF. Suppose that at the initial moment $t_{\rm in}$ the distance between them is equal to $l(t_{\rm in})$, and both are located in the same causally connected region, "under the horizon". Since the size of a causally connected region is determined by the inverse of the Hubble parameter H, at the first moment there holds the inequality

$$l(t_{\rm in}) < H^{-1}(t_{\rm in}).$$

Further on, the distance between the particles grow along with the scale factor $a(t)$

$$l(t) = a(t)r, \tag{7.94}$$

where the coordinate distance r does not change with time, and the Hubble parameter is by definition $H(t) \equiv \dot{a}(t)/a(t)$. Using these relations, it is easy to find the time dependence of the ratio $l(t)/H(t)^{-1}$. This ratio of the distance between the particles and the horizon size as a function of time is the subject of our interest:

$$\frac{l(t)}{H^{-1}(t)} = \dot{a}(t)r. \tag{7.95}$$

In the inflationary period, the scale factor is $a(t) \propto e^{H_0 t}$, while the Hubble parameter is approximately constant, $H(t) \simeq H_0 = \text{const}$.

Eq. (7.95) acquires the form

$$\left(\frac{l(t)}{H(t)^{-1}}\right)_{\text{infl}} \propto \mathrm{e}^{H_0 t},$$

so that the distance between two particles grows exponentially as compared to the horizon size. Even if these particle were born in a causally connected region, there will be an instant t_1 when they will turn out to be in causally disconnected regions, for which $l(t_1) > H^{-1}(t_1)$. It is said that at this moment the distance between the particles "has crossed the horizon".

Thus the previously close particles turn out to be far from each other, in causally disconnected regions, at the expense of the properties of space during inflation.

What happens when the inflation is over? The scale factor behaves as

$$a(t) \propto t^{\beta},$$

where the parameter β is in the interval $0 < \beta < 1$ at any post-inflationary stage [see (7.57), (7.80)]. Let us calculate the ratio of interest using Eq. (7.95):

$$\left(\frac{l(t)}{H(t)^{-1}}\right)_{\text{power}} \propto t^{\beta-1}, \quad 0 < \beta < 1.$$

We see that the horizon size $H(t)^{-1}$ grows more rapidly than the distance between the two particles. At some time t_2 the horizon size becomes equal to the distance between these particles and later exceeds it. The particles are then again in a single causally connected region. All the above-said certainly refers not only to distances between particles but also to any other phenomena which do not strongly affect the metric, e.g., field fluctuations.

7.7 Basic properties of expanding space

7.7.1 The redshift

Here and henceforth the scale factor is normalized so that $a(t_0) = 1$.

Let a source emit a light signal with the wavelength $\lambda(t)$ at a time t, which is received on Earth with the wavelength λ. At emission, the scale factor is equal to $a(t)$, and at the moment of absorption it is naturally equal to unity.

The redshift is defined as

$$z = \frac{\Delta\lambda}{\lambda} = \frac{\lambda - \lambda(t)}{\lambda}. \tag{7.96}$$

Evidently

$$z = \frac{1 - a(t)}{a(t)} \Rightarrow a(t) = \frac{1}{1+z}. \tag{7.97}$$

Let us suppose that the Universe is flat, $k = 0$. The coordinate distance $r(t)$ and the physical distance $R(t)$ to the source are

$$r(t) = \int_t^{t_0} \frac{dt'}{a(t')}, \tag{7.98}$$

$$R(t) = a(t)r(t) = \frac{1}{1+z} r(t). \tag{7.99}$$

We are interested in the physical distance (7.99). To find it, we will need an expression for $a(t)$ or the related $H(z)$. In more detail,

$$R(t) = a(t) \int_t^{t_0} \frac{dt'}{a(t')} = \frac{1}{1+z(t)} \int_t^{t_0} [1 + z(t')]dt'.$$

The substitution $t' \to z$ leads to the expression

$$dz = -\frac{\dot{a}}{a(t')^2} dt' = -\frac{H(t')}{a(t')} dt' = -H(z)[1 + z(t')]dt'.$$

The dependence $H(z)$ is known: according to (7.87),

$$H(z) = H_0 \sqrt{\Omega_r (1+z)^4 + \Omega_m (1+z)^3 + \Omega_\Lambda}.$$

The quantity $\Omega_k = 0$ since we are considering a flat Universe with $k = 0$. Finally, the distance to the object that has emitted the light signal is

$$R(t) = \frac{1}{1+z} \int_0^z \frac{dz'}{H(z')}.$$

And the redshift z is determined from observations by the frequency shift between the emitted and absorbed signals according to (7.96).

Thus, in the expanding Universe, we can find the distance to the emitting object. We can also find the velocity of its separation from the observer due to cosmological expansion:

$$v(t) = \dot{R}(t) = \dot{a}(t)r(t) = H(z)R(t) = \frac{H(z)}{1+z} \int_0^z \frac{dz}{H(z)}.$$

The emission time t is also unambiguously related to the redshift z. This relationship can be easily found for any of the main periods using the expressions (7.78), (7.57), (7.80) and the relation (7.97).

Problem. Observers have detected radiation from a quasar with $z = 6.41$. Find: when and at what distance from the Earth was this light emitted? What is the speed of the quasar measured from the Earth?

Let us note that, apart from the cosmological redshift, another, gravitational redshift is also well known, and it is one of the classical observable effects of relativistic gravity; it is closely related to the time slowing-down near gravitating masses. It has no direct relationship with cosmology but must be taken into account for interpretation of astronomical observations.

7.7.2 The luminosity distance

An important role in observations is played by the notion of luminosity distance d_L. Let an object emit an energy E per unit time in the form of photons or gravitational waves. An observer on Earth detects the energy flux F. Since in the expanding Universe the wavelength is changing as the signal moves from the source to the observer, the seeming luminosity of the object is $L = E/a(t)$. It is simply the effect of photon energy decrease due to increased wavelength. It is natural to define the luminosity distance through the relation

$$F = \frac{L}{4\pi d_L^2}.$$

Then the relationship between the energy emitted by the source and that absorbed by the receiver on Earth at $t = t_0$ is

$$F dt_0 \cdot 4\pi r^2 = E\, dt = L\, a(t)\, dt. \tag{7.100}$$

Let us prove the relationship

$$\frac{dt}{dt_0} = a(t). \tag{7.101}$$

For a signal emitted at t and received at present, t_0, by an observer at the coordinate distance r, we have

$$\int_t^{t_0} \frac{dt'}{a(t')} = r.$$

The same after the time interval dt:

$$\int_{t+dt}^{t_0+dt_0} \frac{dt'}{a(t')} = r.$$

Comparing the two expressions for r, we obtain (7.101) and hence

$$F = \frac{La^2(t)}{4\pi r^2}.$$

Returning to (7.100) and (7.101), we obtain

$$d_L = \sqrt{\frac{L}{4\pi F}} = \sqrt{\frac{L}{4\pi La^2(t)/(4\pi r^2)}} = \frac{r}{a(t)} = \frac{R_{\text{phys}}}{a^2(t)},$$

and finally we obtain a relationship between the luminosity distance and the redshift in the form

$$d_L = \frac{R_{\text{phys}}}{(1+z)^2}. \tag{7.102}$$

7.7.3 The velocity of particles in FRW space-time

The properties of matter, its density and its equation of state, affect the global properties of the space itself and, in particular, its expansion dynamics, i.e., the time dependence of its scale factor. But the opposite is also true: it turns out that the expansion of space diminishes the velocities of particles. Let us prove it, for example, for massive nonrelativistic particles.

Let us first prove the Hubble law for arbitrary times and distances. Let there be two observers located at a coordinate distance r from each other, both at rest in the comoving RF, i.e., $r = \text{const}$. The physical distance varies in time, $R(t) = a(t)r$. Thus the physical velocity of one observer with respect to the other is also variable. Indeed,

$$v = \dot{R} = a(t)r = H(t)a(t)r = H(t)R. \tag{7.103}$$

Evidently, the Hubble parameter $H(t)$ can be considered to be approximately constant only at the present epoch.

Consider the motion of a particle from one observer to the other, located at the coordinate distance dr, under the same conditions, so that $dr = \text{const}$. Physically, according to the Hubble law, the second observer moves away from the first one with the velocity

$$dv_{\text{obs}} = H(t)dR,$$

where dR is the physical distance between them. A particle having the velocity $V(t)$ with respect to the first observer and passing near him at the time $t = t_0$ will reach the second observer in the time

$$dt = dR/V(t).$$

If the space were not expanding, the particle velocity $V(t)$ would be constant. In the expanding Universe it is not the case. In the time interval dt, the particle velocity is measured by the second observer who moves relative to the first one with the velocity dv_{obs}. He will find that the velocity of the particle passing by him is

$$V(t+dt) = V(t) - dv_{\text{obs}} = V(t) - H(t)dR.$$

This equality shows that

$$\frac{dV}{dt} = -H(t)\frac{dR}{dt} = -H(t)V = -\frac{V}{a}\frac{da}{dt}.$$

The last equality is simply substituting the definition of the Hubble parameter. Finally, we have the following trivial equation:

$$dV/V = -da/a,$$

whose solution is

$$V(t) = \text{const}/a(t). \tag{7.104}$$

The constant is, as usual, determined by the initial conditions.

Thus in the expanding Universe the velocity of a free particle decreases with time. The Newton laws hold only in Minkowski space in which $a(t)=1$. Moreover, unlike both Newtonian mechanics and special relativity, there is a distinguished RF in nature: it is the one in which the CMB is isotropic.

Chapter 8

Field dynamics in the inflationary period

The inflationary paradigm has been developed for about 30 years, beginning with the papers [191, 382, 383]. It has allowed for successfully solving the basic cosmological problems, beginning with the earliest stages at which our Universe was formed and ending with the galaxy formation stage. The existence of an inflationary period in the history of our Universe seems inevitable since it explains a great number of observational facts [232, 272]. The first inflationary mechanisms [191, 383] were based on self-consistent equations of the scalar and gravitational fields. They solved in the most economic manner the key problems of the Big Bang theory: the horizon problem, the flatness problem etc. The key point is that, under certain conditions, the interacting gravitational and scalar fields lead to an exponential growth of initially small spatial regions. The scalar "inflaton" field ultimately decays into fermions, thus heating the Universe [275].

Nevertheless, the simplest inflationary models usually require substantial improvement. For instance, their predictions concerning the CMB temperature fluctuations agree with reality only with unnaturally small parameters of the potential for the inflaton field which is responsible for the very opportunity of the inflationary process. On the other hand, its interaction with the matter fields must be sufficient for creating, in the post-inflationary period, the observed baryon and lepton numbers.

A further development of the theory has led to emergence of a large number of inflationary models containing additional fields, such as, for example, hybrid inflation [274] and inflation with pseudo-Nambu-Goldstone field [139]. Most of these models are based on interactions between the classical components of different fields. Meanwhile, their nature contains numerous kinds of fields whose interactions must lead to new phenomena during inflation. Thus, for instance, dissipation effects lower the probabilities of phase transitions. Some inflationary models include as one of their basic elements an interaction between the classical inflaton field with various sorts of particles it has created. This effect is a basis for the model of "warm inflation"; back reaction of the created particles on the dynamics of the inflaton field has been discussed in [141].

8.1 Quadratic inflation

Consider inflation with a quadratic potential as one of the most promising models. The general equations have been presented in the previous chapter. It is well known that the inflaton field, as a number of other scalar fields, emerges in multidimensional theories in a natural way. The shape of the potential can be quite complicated and strongly depends on the theory chosen. However, our interest is in the shape of the potential near its minimum since our Universe was formed at the last stage of inflation, in the process of its completion. But near a minimum, any reasonable potential can be expanded in a Taylor series and approximated by its quadratic term. Thus we choose as an example the simplest, quadratic potential

$$V(\varphi) = \frac{1}{2}m^2\varphi^2. \tag{8.1}$$

Now the general equations of Sec. 7.3 can be given in a specific form. In particular, the set of Eqs. (7.38) and (7.39) is solved analytically. The solution has the form

$$\varphi(t) = \varphi_{\rm in} - \frac{mM_{\rm Pl}}{\sqrt{12\pi}}t. \tag{8.2}$$

Recall that this solution is valid as long as the inequality (7.40) holds. For the quadratic potential it reduces to the inequality

$$\varphi \gg \frac{M_{\rm Pl}}{2\sqrt{3\pi}}. \tag{8.3}$$

It is useful to know the field value at which the slow-rolling conditions are violated, i.e., when $\epsilon \simeq 1$, $\eta \simeq 1$, and therefore inflation terminates. It can be easily found that for the quadratic potential we are considering this happens at

$$\varphi_{\text{end}} \approx \frac{1}{2\sqrt{\pi}} M_{\text{Pl}}. \tag{8.4}$$

Thus the φ field (the inflaton) should be at the Planck scale to make the conditions for an inflationary stage.

Let us estimate the growth of the spatial size of the filed fluctuation. To do that, one should find the number of e-folds N, since the size of the Universe grows by a factor of e^N. This number is easily obtained from Eq. (7.47) in our special case of a quadratic potential:

$$N = \frac{2\pi}{M_{\text{Pl}}^2}(\varphi_{\text{in}}^2 - \varphi_{\text{end}}^2). \tag{8.5}$$

It is known from observations that $N \approx 60$. Now we can estimate the inflaton value at the moment when the modern horizon was created, i.e., at creation of the spatial region that we call the observable Universe:

$$\varphi_{\text{in}} \approx M_{\text{Pl}} \sqrt{\frac{N}{2\pi}} \sim 3M_{\text{Pl}}. \tag{8.6}$$

It has been taken into account here that $\varphi_{\text{end}} \ll \varphi_{\text{in}}$ and that the number of e-folds known from observations is $N \approx 60$. It is now easy to estimate the value of the potential (8.1) at the creation of the modern horizon:

$$V(\varphi_{\text{in}}) = \frac{1}{2} m^2 \varphi_{\text{in}}^2 \approx 4.5 m^2 M_{\text{Pl}}^2 \approx 4.5 \cdot 10^{-12} M_{\text{Pl}}^4.$$

The last equality has been obtained for the usually-supposed value of the inflaton mass, $m \approx 10^{-6} M_{\text{Pl}}$. Thus the modern horizon of our Universe appeared at energies much smaller than the Planck one.

Let us see what is the size of the horizon that appeared at the Planck energy scale, at $V \sim M_{\text{Pl}}^4$. In this case, the initial field value is $\varphi_{\text{Planck}} \sim M_{\text{Pl}}^2/m \sim 10^6 M_{\text{Pl}}$. Substituting this value to the expression (8.5), we obtain

$$N \sim 2\pi \cdot 10^{12}.$$

So the linear size of the spatial region that emerged at the Planck energy scale has increased by a factor of $\exp(2\pi \cdot 10^{12})$. Its initial size was of the Planck order. Right after the end of inflation, the spatial size of such a region is

$$L \sim \exp(2\pi \cdot 10^{12}) \cdot 10^{-33}\,\text{cm} \sim 10^{2\cdot 10^{12}}\,\text{cm}.$$

8.2 Quantum fluctuations during inflation

A brief analysis

As is known, the classical evolution of systems is slightly changed by quantum fluctuations. In Minkowski space their role is insignificant in many cases because the quantum corrections are proportional to a small parameter, the Planck constant $\hbar$. In addition, according to Heisenberg's uncertainty principle, the larger a fluctuation, the smaller its existence time. The situation is radically different at the inflationary stage which is approximately described in terms of de Sitter space. The main property of inflation is that any spatial inhomogeneity is stretched along with the expansion of space. Field fluctuations are also inhomogeneities, and their behaviour is quite unlike the behaviour of fluctuations in Minkowski space. In de Sitter space, their size rapidly grows and exceeds the horizon size. This process resembles particle creation in a strong field. After creation, virtual particles fly apart instead of annihilating, as happens in Minkowski space. The additional energy is created by the work of the external field, the gravitational field in our case.

The hypothetic field that makes the inflationary stage possible is generally assumed to be scalar. One of the Einstein equations where this scalar field φ is a source has the form

$$\ddot{\varphi} + 3H\dot{\varphi} - H^{-2}\,\mathrm{e}^{-2Ht}\Delta\varphi + V'(\varphi) = 0. \tag{8.7}$$

Eq. (8.7) follows from Eqs. (7.37) and (7.4) after simple calculations. Here, we have neglected the weak change of the Hubble parameter $H = \dot{a}/a$ at the inflationary stage, so that the scale factor is $a(t) = H^{-1}\,\mathrm{e}^{Ht}$.

Now we divide the field into "classical" (Φ) and "quantum" (q) parts:

$$\varphi(\mathbf{x}, t) = \Phi(\mathbf{x}, t) + q(\mathbf{x}, t). \tag{8.8}$$

The term $q(\mathbf{x}, t) \equiv \delta\phi(\mathbf{x}, t)$, representing small fluctuations, must be quantized, and to do so, we employ, as usual, the Fourier expansion

$$\delta\varphi(x, t) = \int \frac{d^3k}{(2\pi)^3}\left[a_k f_k(t)\,\mathrm{e}^{ikx} + \mathrm{h.c.}\right],$$

where a_k is the annihilation operator for the mode k. Note that in this analysis we are using the comoving RF with a dimensionless coordinate x, and the wave vector k is also dimensionless. From the basic Eq. (8.7), putting $H \simeq \text{const}$, we obtain

$$\ddot{f}_k + 3H\dot{f}_k + \frac{k^2}{(H\,\mathrm{e}^{Ht})^2} f_k + V''(\Phi) f_k = 0. \tag{8.9}$$

During inflation the field is changing slowly, therefore we neglect the last term, which allows for finding a suitable solution to this equation

$$f_k(t) = \frac{H}{\sqrt{2k^3}}(i + k\,\mathrm{e}^{-Ht})\,\mathrm{e}^{ik\,\mathrm{e}^{-Ht}}. \tag{8.10}$$

The behavior of the solution is extremely curious. At large wave numbers it rapidly oscillates while at small wave numbers it tends to a constant. A critical wave number that separates the two regimes is $k_H = a(t)H(t) = \mathrm{e}^{Ht}$. This wave number is inversely proportional to the characteristic size of the region, which is in this case a fluctuation emerging at a certain instant at the inflationary stage.

The following picture is appearing. Quantum effects create a fluctuation of the φ field whose spatial size, l, is naturally smaller than the horizon size because a fluctuation can only appear in a causally connected region. The condition $l < 1$ means $k > 1$ (recall that we are using dimensionless coordinates and momenta).

Let us suppose that there initially emerged a small-scale field configuration with the size $l_0 \ll 1$ and the characteristic momentum $k_0 \gg 1$. The function f_k describing this configuration, (8.10), rapidly oscillates. The limiting momentum $k_H(t) = \mathrm{e}^{Ht}$ grows exponentially rapidly, and there must occur such an instant, t_H, that the characteristic momentum of the system will be comparable with the limiting momentum, $k_0 \sim k_H(t_H)$. Beginning with this time, the function f_k stops to oscillate and tends to a constant. It is clear that the values of this constant are randomly distributed. The distribution law is considered in the next section.

So far we have been dealing with comoving coordinates. Let us now look at these processes from the viewpoint of physical distances $R_{\text{phys}} = a(t)r = H^{-1}\,\mathrm{e}^{Ht}r$ and the corresponding physical momenta $P_{\text{phys}} = p/a(t) = H\mathrm{e}^{-Ht}k$. The limiting physical momentum at which the magnitude of a certain fluctuation is "frozen" is $P^H_{\text{phys}} = H\mathrm{e}^{-Ht}k_H = H$. The characteristic size is then $L_{\text{phys}} = 1/P^H_{\text{phys}} = 1/H$. Thus a fluctuation emerging at the inflationary stage due to quantum effects develops in the following way: the field value fluctuates very rapidly while its spatial size exponentially grows until it becomes the inverse value of the Hubble parameter.

A detailed consideration

Let us return to the expansion (8.8). The classical part of Φ is associated with smooth and slow changes of the field. The following expansion on the basis of the Fourier transformation is often used:

$$\varphi(\mathbf{x},t) = \int \frac{d^3\mathbf{k}}{(2\pi)^{3/2}}[a_{\mathbf{k}}(t)\,\mathrm{e}^{-i(\mathbf{kx})} + a^{\dagger}_{\mathbf{k}}(t)\,\mathrm{e}^{i(\mathbf{kx})}];$$

$$\Phi(\mathbf{x},t) = \int_{|\mathbf{k}|<k^*} \frac{d^3\mathbf{k}}{(2\pi)^{3/2}}[a_{\mathbf{k}}(t)\,\mathrm{e}^{-i(\mathbf{kx})} + a^{\dagger}_{\mathbf{k}}(t)\,\mathrm{e}^{i(\mathbf{kx})}],$$

$$q(\mathbf{x},t) = \int_{|\mathbf{k}|>k^*} \frac{d^3\mathbf{k}}{(2\pi)^{3/2}}[a_{\mathbf{k}}(t)\,\mathrm{e}^{-i(\mathbf{kx})} + a^{\dagger}_{\mathbf{k}}(t)\,\mathrm{e}^{i(\mathbf{kx})}], \tag{8.11}$$

where the critical momentum k^* separates the “rapid” subsystem from the “slow”, classical one. This choice is not unique and depends on the problem under consideration. In Minkowski space, the most convenient is the choice of basis functions of the form $a_{\mathbf{k}}(t) \sim \mathrm{e}^{ik_0 t}$, which are solutions of the d’Alembert equation

$$\Box g(\mathbf{x},t) = 0.$$

In de Sitter space it is also convenient to choose the solutions of this equation as basis functions. After the Fourier transformation

$$\tilde{g}_{\mathbf{p}}(t) = \int d^3x\,\mathrm{e}^{i(\mathbf{px})} g(\mathbf{x},t),$$

the equation acquires the form

$$\frac{\partial^2 \tilde{g}_{\mathbf{p}}(t)}{\partial t^2} + 3H\frac{\partial \tilde{g}_{\mathbf{p}}(t)}{\partial t} + (H\mathbf{p})^2\,\mathrm{e}^{-2Ht}\tilde{g}_{\mathbf{p}}(t) = 0, \tag{8.12}$$

where H is the Hubble parameter. Let us note that the momentum $\mathbf{p}$ is dimensionless as well as the comoving coordinate $\mathbf{x}$. One often uses the momentum $\mathbf{P} \equiv H\mathbf{p}$, having the correct dimensionality. Then the set of solutions of Eq. (8.12) is expressed in terms of Hankel functions:

$$H^{(2)}_{3/2}(y) = [H^{(1)}_{3/2}(y)]^* = -\sqrt{\frac{2}{\pi y}}\,\mathrm{e}^{-iy}\left(1 + \frac{1}{iy}\right).$$

The solution can be presented in the form

$$\tilde{g}_{\mathbf{p}}(t) = \frac{\sqrt{\pi}}{2}H\eta^{3/2}[c_1(p)H^{(1)}_{3/2}(\eta P) + c_2(p)H^{(2)}_{3/2}(\eta P)].$$

Here, the so-called conformal time has been introduced:

$$\eta \equiv -H^{-1}\,\mathrm{e}^{-Ht}. \tag{8.13}$$

This quantity simplifies the equations when analyzing processes in de Sitter space. The two constants are determined from additional conditions. More precisely, it is natural to suppose that at distances much smaller than the horizon size, i.e., as $p \to \infty$, the solution must coincide with that in Minkowski space. It is possible if $c_1 = 0,\ c_2 = -1$. Thus the set of orthonormalized solutions has the form [272]

$$\tilde{g}_{\mathbf{p}}(t) = \frac{iH}{P^{3/2}\sqrt{2}}\left(1 + \frac{P}{iH}\,\mathrm{e}^{-Ht}\right)\exp\left(\frac{iP}{H}\,\mathrm{e}^{-Ht}\right). \tag{8.14}$$

An analysis of the behaviour of $\tilde{g}_{\mathbf{p}}(t)$ indicates the existence of the threshold value of the momentum,

$$P^*(t) \equiv H\,\mathrm{e}^{Ht}. \tag{8.15}$$

At each time t, the solution oscillates at $P \gg P^*$ and tends to a constant at $P \ll P^*$:

$$\tilde{g}_{\mathbf{p}}(t) \simeq \frac{iH}{P^{3/2}\sqrt{2}}, \quad P \ll P^*. \tag{8.16}$$

To make clear the physical meaning of the result obtained, let us pass on to physical distances, $\mathbf{R}_{\text{phys}} = a(t)\mathbf{r}$, and momenta. Evidently, the physical momentum $\mathbf{P}_{\text{phys}}$ is connected with the coordinate one, $\mathbf{p}$, in the following way: $\mathbf{P}_{\text{phys}} = \mathbf{p}/a(t) = \mathbf{P}/(a(t)H)$. At the inflationary stage, the scale factor is related to the Hubble parameter, $a(t) \simeq H^{-1}\,\mathrm{e}^{Ht}$. Consequently, the threshold value of the physical momentum P^*_{phys} is

$$P^*_{\text{phys}} \equiv \frac{P^*}{a(t)H} = H.$$

The characteristic value of a fluctuation is $L_{\text{phys}} \sim P^{-1}_{\text{phys}}$. If this size is larger than the horizon size H^{-1}, the magnitude of the fluctuation tends to a constant. Note that the physical size L_{phys} of the fluctuation grows exponentially:

$$L_{\text{phys}} \sim P^{-1}_{\text{phys}} = \frac{a(t)H}{P} = \frac{1}{P}\,\mathrm{e}^{Ht}. \tag{8.17}$$

This behaviour is just characteristic of de Sitter space. In Minkowski space, the lifetime of a quantum fluctuation is of the order $1/\Delta E$ (ΔE is the fluctuation energy). In de Sitter space, a quantum fluctuation, having been born, increases its size exponentially according to the expression (8.17). At the same time, its magnitude is determined by the expression (8.16).

Now we can return to the problem of splitting the scalar field φ into classical and quantum parts. The quantum part, see (8.11), can be written in the form

$$q(\mathbf{x},t) \equiv \int \frac{d^3\mathbf{p}}{(2\pi)^{3/2}} W(P,t) \left[\hat{a}_{\mathbf{p}}\tilde{g}_{\mathbf{p}}(t)\,\mathrm{e}^{-i(\mathbf{px})} + \hat{a}^{\dagger}_{\mathbf{p}}\tilde{g}^{*}_{\mathbf{p}}(t)\,\mathrm{e}^{i(\mathbf{px})}\right]. \qquad (8.18)$$

Here we have introduced the creation and annihilation operators $\hat{a}^{\dagger}_{\mathbf{p}}$ and $\hat{a}_{\mathbf{p}}$, according to the standard manner of field quantization. Instead of roughly cutting off the momentum by the condition $P > P^*$, we use the function $W(P,t)$ with the properties $W(P \to 0,t) \to 0$, $W(P \to \infty,t) \to 1$ [385]. This can be, e.g., the function

$$W(P,t) = \theta(P - \varepsilon H\,\mathrm{e}^{Ht}), \quad \varepsilon \ll 1. \qquad (8.19)$$

As will be seen later, the final result does not depend on the small but otherwise arbitrary parameter ε. Substituting the expressions (8.8) and (8.18) to Eq. (8.7), we obtain

$$\begin{aligned} &\frac{\partial\Phi}{\partial t} - \frac{1}{3H}\left[\mathrm{e}^{-2Ht}\Delta\Phi - \frac{\partial V(\Phi)}{\partial\Phi}\right] = y(\mathbf{x},t),\\ &y(\mathbf{x},t) \equiv \left(\frac{1}{3H}\frac{\partial^2}{\partial t^2} - \frac{\partial}{\partial t} + \frac{1}{3H}\,\mathrm{e}^{-2Ht}\Delta\right) q(\mathbf{x},t). \end{aligned} \qquad (8.20)$$

In this equation we have made an approximation: taking into account the slow change of the field during inflation, we omit the second-order derivative and neglect the higher powers of the function $y(\mathbf{x},t)$. Eq. (8.20) describes the dynamics of the Φ field subject to the random "force" y. The latter is assumed to be small, so that we can seek the solution in the form [342]

$$\Phi = \Phi_{\mathrm{det}} + \phi. \qquad (8.21)$$

The main part of classical field Φ_{det} obeys the equation

$$\frac{\partial\Phi_{\mathrm{det}}}{\partial t} - \frac{1}{3H}\left[\mathrm{e}^{-2Ht}\Delta\Phi_{\mathrm{det}} - \frac{\partial V(\Phi_{\mathrm{det}})}{\partial\Phi_{\mathrm{det}}}\right] = 0, \qquad (8.22)$$

whereas the random part ϕ entirely depends on quantum fluctuations according to the linear equation

$$\frac{\partial\phi}{\partial t} - \frac{1}{3H}[\mathrm{e}^{-2Ht}\Delta\phi - V''(\Phi_{\mathrm{det}})\phi] = y(\mathbf{x},t). \tag{8.23}$$

(We consider the limit $\Phi_{\mathrm{det}} \gg \phi$, which is justified if the random "force" $y(\mathbf{x},t)$ is small). Calculating the random "force" appearing due to quantum fluctuations, we obtain (see details in [342])

$$\begin{aligned} y(\mathbf{x},t) &\equiv \left(\frac{1}{3H}\frac{\partial^2}{\partial t^2} - \frac{\partial}{\partial t} + \frac{1}{3H}\,\mathrm{e}^{-2Ht}\Delta\right) q(\mathbf{x},t) \\ &= \left(\frac{1}{3H}\frac{\partial^2}{\partial t^2} - \frac{\partial}{\partial t} + \frac{1}{3H}\,\mathrm{e}^{-2Ht}\Delta\right) \int \frac{d^3\mathbf{p}}{(2\pi)^{3/2}}\theta\left(P - \varepsilon H\,\mathrm{e}^{Ht}\right) \\ &\quad \times [\hat{a}_{\mathbf{p}}\tilde{g}_{\mathbf{p}}(t)\,\mathrm{e}^{-i(\mathbf{px})} + \hat{a}^{\dagger}_{\mathbf{p}}\tilde{g}^{*}_{\mathbf{p}}(t)\,\mathrm{e}^{i(\mathbf{px})}] \\ &\simeq \frac{H^3\varepsilon}{\sqrt{2}}\,\mathrm{e}^{Ht}\int \frac{d^3p}{(2\pi p)^{3/2}}\delta(P - \varepsilon H\,\mathrm{e}^{Ht})[\hat{a}_{\mathbf{p}}\,\mathrm{e}^{-i(\mathbf{px})} - \hat{a}^{\dagger}_{\mathbf{p}}\,\mathrm{e}^{i(\mathbf{px})}]. \end{aligned} \tag{8.24}$$

Here we have used Eq. (8.12), which substantially simplifies the calculations; omitted is the second-order derivative, proportional to the small parameter ε^2. We also have $P = \sqrt{\mathbf{P}^2} = \sqrt{(\mathbf{p}H)^2}$. In the last line we use the approximation (8.16). The validity of this assumption follows from smallness of the momentum that enters into the argument of the delta function.

Consider the correlator

$$D(\mathbf{x},t,\mathbf{x}',t') \equiv \langle 0\,|y(\mathbf{x},t), y(\mathbf{x}',t')|\,0\rangle\,.$$

Using the expression (8.24) obtained above and the properties of the creation and annihilation operators $a^{\dagger}$ and a, we can obtain an analytic expression for this quantity:

$$D(\mathbf{x},t,\mathbf{x}',t') = \frac{H^3}{4\pi^2}\delta(t-t')\frac{\sin\varepsilon|\mathbf{x}-\mathbf{x}'|\,\mathrm{e}^{Ht}}{\varepsilon|\mathbf{x}-\mathbf{x}'|\,\mathrm{e}^{Ht}}, \qquad \varepsilon \ll 1. \tag{8.25}$$

Homogeneous fluctuations

According to the expression (8.25), the correlator $D(\mathbf{x},t,\mathbf{x}',t')$ turns out to be an abrupt function of the distance $|\mathbf{x}-\mathbf{x}'|$. For the same reasons, one can neglect the spatial derivatives in Eq. (8.23), which greatly simplifies the equations. So there is a reason to consider a homogeneous distribution $\Phi = \Phi(t)$ obeying the simple equation

$$\frac{\partial \Phi}{\partial t} + \frac{1}{3H}\frac{\partial V(\Phi)}{\partial \Phi} = 0, \tag{8.26}$$

$$\frac{\partial \phi}{\partial t} + \frac{m^2}{3H}\phi = y(t). \tag{8.27}$$

Here we have introduced the notation

$$m^2 \equiv V''(\Phi_{\rm det}).$$

For the quadratic potential

$$V(\phi) = V_0 + \frac{1}{2}m^2\phi^2,$$

this quantity is a constant, while in the general case of an arbitrary potential it is almost constant during inflation. The correlator (8.25) of the random function $y(t)$ in the limit $\varepsilon \ll 1$ can be approximated as follows:

$$\langle y(t_1)y(t_2)\rangle = D(\mathbf{x}, t, \mathbf{x}, t') = \frac{H}{4\pi^2}\delta(t_1 - t_2).$$

The emergence of a delta function in the right-hand side of the equation means the Gaussian distribution law of the random function $y(t)$ with the energy

$$W(y) = \text{const} \cdot \exp\left[-\frac{1}{2\sigma^2}\int y(t)^2 dt\right], \quad \sigma = \frac{H^{3/2}}{2\pi}.$$

Due to a linear relation between the functions ϕ and $y(t)$, see (8.27), their probability distributions are proportional to each other. It means that the probability of finding the value of the function $\phi(t)$ inside a certain small interval is equal to (see [162])

$$dP(\phi) = \text{const} \cdot \mathcal{D}\phi \exp\left[-\frac{1}{2\sigma^2}\int \left[\frac{\partial \phi}{\partial t} + \frac{m^2}{3H}\phi\right]^2 dt\right].$$

The measure has the form $\mathcal{D}\phi \equiv \prod_{i=1}^{N} d\phi(t_i)$, $N \to \infty$.

It is now easy to obtain the probability of finding the field value ϕ_2 at the moment t_2 under the condition that the quantity ϕ_1 at t_1 is known. Evidently, it is necessary to integrate over all field values in the interval

(t_1, t_2), except for the boundary points $\phi_1 \equiv \phi(t_1)$, $\phi_2 \equiv \phi(t_2)$, which leads to the expression

$$dP(\phi_2, t_2; \phi_1, t_1) = \text{const} \cdot d\phi_2 \int_{\phi_1}^{\phi_2} \mathcal{D}\phi \exp\left[-\frac{1}{2\sigma^2}\int_{t_1}^{t_2}\left[\frac{\partial\phi}{\partial t} + \frac{m^2}{3H}\phi\right]^2 dt\right]. \tag{8.28}$$

The unknown constant is found from the normalization condition

$$\int_{-\infty}^{\infty} d\phi_2 P(\phi_2, t_2;\ \phi_1, t_1) = 1.$$

The functional integral (8.28) can be calculated exactly [162] using the saddle point method, which implies a search for an extremum of the action

$$\ddot{\phi} - \mu^2\phi = 0, \quad \mu \equiv \frac{m^2}{3H},$$

with the boundary condition

$$\phi(t_1) = \phi_1, \quad \phi(t_2) = \phi_2.$$

A solution of this equation reads

$$\begin{aligned}\phi(t) &= A\,\mathrm{e}^{\mu t} + B\,\mathrm{e}^{-\mu t},\\ A &= \frac{\phi_2 - \phi_1\,\mathrm{e}^{-\mu T}}{2\sinh(\mu T)}, \quad B = \frac{-\phi_2 + \phi_1\,\mathrm{e}^{\mu T}}{2\sinh(\mu T)}; \quad T = t_2 - t_1.\end{aligned}$$

Substituting it into the integral in the exponential in the expression (8.28), we find the required probability as

$$\begin{aligned}&dP(\phi_2, t_1 + T; \phi_1, t_1) = d\phi_2 \cdot \sqrt{\frac{r}{\pi}}\exp[-r(\phi_2 - \phi_1\,\mathrm{e}^{-\mu T})^2],\\ &r \equiv \frac{\mu}{\sigma^2}\frac{1}{1 - \mathrm{e}^{-2\mu T}},\\ &\mu = \frac{m^2}{3H} \simeq \text{const}, \quad \sigma = \frac{H^{3/2}}{2\pi} \simeq \text{const}.\end{aligned} \tag{8.29}$$

In the massless field limit we arrive at a simpler expression

$$dP(\phi_2, t_1 + T; \phi_1, t_1) = d\phi_2\sqrt{\frac{2\pi}{H^3 T}}\exp\left[-\frac{2\pi^2}{H^3 T}(\phi_2 - \phi_1)^2\right]. \tag{8.30}$$

This formula is widely used in different inflationary scenarios where the field evolution should be small and therefore the second-order derivative of the potential and hence the mass can be neglected.

The field evolution picture looks as follows. The field is assumed to consist of two components, the classical (deterministic) one and the quantum one, see the expression (8.21). The classical part $\Phi_{\rm det}$ obeys the equations of motion (8.26) with a random "force" y. As shown above, the influence of this force is included in the random component ϕ and has the probability density (8.29). It is easy to calculate $\langle \phi^2 \rangle$ for estimating the time-depending deflections from the classical value $\Phi_{\rm det}$. Using the expression for the probability (8.29), we obtain

$$\begin{aligned} \langle \phi(t)^2 \rangle &= \int_{-\infty}^{\infty} \phi^2 \, dP(\phi, t_1 + t; \phi_1, t_1) \\ &= \frac{1}{2r} = \frac{\sigma^2}{2\mu} \left(1 - \mathrm{e}^{2\mu t}\right). \end{aligned} \tag{8.31}$$

This expression can be substantially simplified in the case of a massless field. As already said, the field can be regarded as approximately massless during inflation, when the condition $m \ll H$ holds. Expanding the exponential in the Eq. (8.31), we obtain

$$\sqrt{\langle \phi(t)^2 \rangle} = \sigma = \frac{H}{2\pi}\sqrt{Ht}. \tag{8.32}$$

In terms of e-folds, $N \equiv Ht$, we obtain a formula to be widely used in what follows:

$$\sqrt{\langle \phi(t)^2 \rangle} = \sigma = \frac{H}{2\pi}\sqrt{N}. \tag{8.33}$$

In particular, one can conclude that fluctuations with the magnitude $\sim H/2\pi$ are formed in the time interval $t \sim H^{-1}$ $(N = 1)$. The expressions (8.15), (8.17) indicate that its spatial size becomes fixed at the scale $L_{\rm phys} \sim H^{-1}$.

The spatial size of the fluctuation can also be estimated. To that end, let us note that the correlator (8.25) is not small at $t \sim H^{-1}$ if the comoving distance $|\mathbf{x} - \mathbf{x}'| \lesssim 1$. It means that those fluctuations with comoving size of the order of unity are significant. Their size in physical coordinates grows in the usual way,

$$L_{\rm fluct} = a(t)|\mathbf{x} - \mathbf{x}'| \approx H^{-1}\, \mathrm{e}^{Ht}, \tag{8.34}$$

and is equal to H^{-1} at $t \sim H^{-1}$. This remark is very important for what follows. By the way, the results (8.33) and (8.34) can be easily reproduced even in Minkowski space. Indeed, choosing the Lagrangian for a massless scalar field, let us make order-of-magnitude estimates for the action

$$S = \int d^4x\sqrt{-g}\frac{1}{2}g^{\mu\nu}\partial_\mu\varphi\partial_\nu\varphi.$$

As a result, we obtain

$$S \sim \frac{(\Delta\varphi)^2}{H^{-2}}H^{-4}.$$

Here $\Delta\varphi$ is the magnitude of a fluctuation emerging by the time $t \sim H^{-1}$. It has also been taken into account that the fluctuation size is of the order H^{-1}, which is true for a massless field propagating with the speed of light. The probability of such a fluctuation is of the order of unity under the condition that the action $S \sim 1$. Accordingly, we get the estimate

$$\Delta\varphi \sim H,$$

which coincides by order of magnitude with the result (8.33) obtained more rigourously.

There also exists another way of studying fluctuations, based on the fact that Eq. (8.27) is nothing else than the Langevin equation. Then one can at once make the statement that the probability density satisfies the Fokker–Planck equation

$$\frac{\partial P}{\partial t} = \frac{H^3}{8\pi^2}\frac{\partial^2 P}{\partial\phi^2} + \frac{m^2}{3H^2}\frac{\partial^2(\phi P)}{\partial\phi^2}. \tag{8.35}$$

The expression (8.29) for the probability obtained above is a solution of this equation ($dP = Pd\phi$ in our notations).

8.3 Hybrid inflation

The main shortcoming of the simplest model with a quadratic potential is the small value of the inflaton mass as compared to the Planck mass, $m/M_{\rm Pl} \simeq 10^{-6}$. There are different opportunities to improve the situation, such as, for instance, introduction of one more hypothetic scalar field. One of the most well-known attempts of this kind is hybrid inflation [274, 281].

The potential of hybrid inflation depends on two scalar fields σ and χ and is usually written as

$$V(\chi,\sigma) = \varkappa^2\left(M^2 - \frac{\chi^2}{4}\right)^2 + \frac{\lambda^2}{4}\chi^2\sigma^2 + \frac{1}{2}m^2\sigma^2, \tag{8.36}$$

where $\varkappa$, λ and m are parameters of the theory.

Such a form allows for obtaining in a natural way a slow rolling along the valley $\chi = 0$, $\sigma > \sigma_c$ and rapid fluctuations of both fields after the end of inflation. The slow rolling is necessary for carrying out the inflationary stage while rapid oscillations are necessary for effective production matter particles and heating the Universe.

The inflation is going on until the fields reach the critical point

$$\sigma_c = \sqrt{2}\frac{\varkappa}{\lambda}M.$$

After the field σ passes this value, the motion along the line $\chi = 0$, $\sigma < \sigma_c$ becomes unstable, and the field χ rapidly reaches one of the minima $\chi_{\pm} = \pm 2M$, $\sigma = 0$. The fate of the Universe depends on the initial data, χ_{in}, σ_{in}.

Nevertheless, even this well-studied picture has a serious shortcoming. During the inflationary stage, when the fields σ and χ "move" along the line $\chi = 0$, the space splits into a number of causally disconnected regions. Due to quantum fluctuations, the scalar fields in these regions slightly differ from each other. Evidently, the overwhelming majority of regions contain the field $\chi \neq 0$. So, at the end of inflation, after the field σ reaches its critical value σ_c, there emerges an enormous number (of the order of $\exp(3 \cdot N_U) \sim 10^{78}$) of domains. In half of them, those containing the field $\chi < 0$, the rolling proceeds to the minimum $\chi_- = -2M$, whereas in the other half it goes to the minimum $\chi_+ = +2M$. After inflation, there appears a Universe consisting of chaotically distributed domains with the field values χ_+ or χ_- inside. Adjacent domains are separated by domain walls because in moving from χ_+ to χ_- it is necessary to visit the point $(\chi = 0, \sigma = 0)$, representing an extremum of the potential. Such a period of "wall domination" in the Universe evolution is unacceptable because it makes the existence of the Universe in its present form impossible. Accordingly, the assumed motion "on the average" along the line $\chi = 0$ is also excluded. It is a very strong restriction imposed on the model of hybrid inflation by taking into account quantum fluctuations. But it can be circumvented both in the framework of this model and in its modifications. Addition of a small slope to the potential substantially eases the situation. A change in the shape of the potential leads to a nontrivial scalar field dynamics and, in particular, to a new mechanism of a transition from the inflationary stage to usual cosmological expansion in FRW space-time with effective production of matter particles. A formation mechanism of massive primordial BHs [354] also exists in the framework of hybrid inflation.

8.4 Influence of massive fields on the process of inflation

In this section we will discuss the influence of quantum fluctuations on the classical evolution of the inflaton field. Our purpose will be a study of the back reaction of the cloud of virtual particles created by the inflaton field on the evolution of the field itself. Let us stress that here we are only considering interaction with virtual particles which are in no way related to the ambient temperature. It is known that such a cloud, possessing inertial properties, can appreciably slow down the system evolution. This phenomenon has been well studied in solid state physics (the polaron effect [259]). The whole analysis is conducted in Minkowski space, i.e., without taking into account gravitational effects.

The inertial properties of the virtual cloud depend not so much on the shape of the interaction potential or on the kinds of particles as on the value of the coupling constant and on the mass of the virtual particles, because it is these parameters that determine the richness of space in particles of different kinds. It is also clear that inclusion of new fields only strengthens the "braking" effect in the classical field evolution. Therefore in what follows we consider the simplest form of interaction allowing for obtaining analytical results. Specifically, we consider, apart from the inflaton field φ, an additional scalar field χ and write the action in the form

$$S = \int d^4x\sqrt{-g}\left[\frac{1}{2}\varphi_{,\mu}\varphi^{,\mu} - V(\varphi) + \frac{1}{2}\chi_{,\mu}\chi^{,\mu} - \frac{1}{2}m_\chi^2\chi^2 - \varkappa\chi u(\varphi)\right]. \tag{8.37}$$

Here $u(\varphi)$ is a polynomial with respect to φ with the power not higher than three for renormalizable theories. The linear term in the additional field χ is only needed in the interaction for finally obtaining observable analytical results, valid for an arbitrary coupling constant $\varkappa$. An interaction of this kind emerges in supersymmetric theories and is discussed in hybrid inflation models [281]. In the paper [141], such a form of interaction is used in a study of the back reaction of the created particles on the classical motion of the inflaton field. As indicated above, the inclusion of real opportunities, e.g., with fermions, will only enhance the braking effect. Since our interest is in the interaction with virtual χ-particles, consider the transition amplitude

$$A(\varphi_i,\, \chi_i = 0;\, \varphi_f,\, \chi_f = 0) = \int_{\varphi_i}^{\varphi_f} D\varphi \int_0^0 D\chi \exp[iS]. \tag{8.38}$$

We here assume that the χ field is sufficiently massive and, by the relevant time, is located near a minimum of its effective potential. Integrating in the field variable χ, we obtain an amplitude of the form (8.38) with the effective action

$$S_{\text{eff}} = \int d^4x\sqrt{-g}\left[\frac{1}{2}\varphi_{,\mu}\varphi^{,\mu} - V(\varphi)\right] + \frac{\varkappa^2}{2}\int d^4x\int d^4x'\sqrt{-g}u(\varphi(x))G(x,x')u(\varphi(x')), \quad (8.39)$$

where $G(x,x')$ is the Green fluctuation of the scalar field χ. This expression is exact but, due to the presence of a nonlocal term, which is physically interpreted as a contribution from a cloud of virtual χ-particles, exact analytical transformations are impossible. Therefore, as in the method of effective action [214], we expand the nonlocal term in (8.39) in powers of $x-x'$. Then the first several terms renormalize the initial parameters of the potential and the field variable φ (the wave function). Since the parameters of the potential are determined by the physical conditions at some energy scale, we will be interested in new terms which cannot be incorporated in renormalizations.

An explicit expansion of the expression (8.39) in a power series in $x-x'$ requires integration of the Green function, which, if we take into account gravitational effects, is a hard problem. Instead, let us use the Green function equation [38] written as

$$G(x,x') = \frac{1}{m_\chi^2}\delta(x-x') - \frac{1}{m_\chi^2}\frac{1}{\sqrt{-g}}\partial_\mu\sqrt{-g}\partial^\mu G(x,x'). \quad (8.40)$$

Substituting this expression into (8.39) and neglecting higher derivatives in φ, we obtain

$$S_{\text{eff}} = \int d^4x\sqrt{-g}\left[\frac{1}{2}\varphi_{,\mu}\varphi^{,\mu} - V_{\text{ren}}(\varphi) - \frac{1}{\sqrt{-g}}\frac{\varkappa^2}{2m_\chi^4}u(\varphi)\partial_\mu\sqrt{-g}\partial^\mu u(\varphi)\right], \quad (8.41)$$

with the corresponding equation of motion

$$\partial_\mu\sqrt{-g}\partial^\mu\varphi + \sqrt{-g}V'_{\text{ren}}(\varphi) + \frac{\alpha^2}{m_\chi^2}u'_\varphi(\varphi)\partial_\mu\sqrt{-g}\partial^\mu u(\varphi) = 0, \quad (8.42)$$

and energy density

$$E = T_{00} = \frac{1}{2}(\partial_0\varphi)^2 + V_{\rm ren}(\varphi) + \frac{\alpha^2}{2m_\chi^2}(\partial_0 u)^2, \tag{8.43}$$

where we have introduced the dimensionless parameter $\alpha \equiv \varkappa/m_\chi$.

The renormalized effective potential $V_{\rm ren}(\varphi)$ contains a sum of contributions from interactions with all possible fields. A contribution from interaction with the χ field is easily found explicitly:

$$V_{\rm ren}(\varphi) = V(\varphi) + \delta V_\varphi = V(\varphi) - (\alpha^2/2)u(\varphi)^2.$$

A shortcoming of the first model of chaotic inflation with the potential $\lambda_{\rm ren}\varphi^4$ is the small value of the coupling constant, $\lambda_{\rm ren}(\sim 10^{-13})$, at which there is agreement with the observational data. It means that all terms in the expression for $\lambda_{\rm ren}$, including δV_χ, should cancel with high accuracy. It is further proved that taking into account a renormalization of the kinetic term makes it possible, in particular, to substantially weaken the restrictions on the parameters of the theory due to observational data.

The terms similar to the last term in the expression (8.41) emerge at renormalization of any theory and are usually interpreted as "quantum corrections to the parameters of the theory depending on the value of the field itself" [214]. In weak fields, a contribution from this term is negligibly small. But at the inflationary stage, at large field values, the last term in Eq. (8.41) can be quite significant.

The classical Eq. (8.42) can also be obtained in another way, using the classical equations that minimize the initial action (8.37),

$$\begin{aligned} \frac{1}{\sqrt{-g}}\partial_\mu\left(\sqrt{-g}\partial^\mu\chi\right) + m_\chi^2\chi + \varkappa\varphi^2 &= 0, \\ \frac{1}{\sqrt{-g}}\partial_\mu\left(\sqrt{-g}\partial^\mu\varphi\right) + V'(\varphi) + 2\varkappa\varphi\chi &= 0. \end{aligned} \tag{8.44}$$

We here consider the case $u(\varphi) = \varphi^2$. The first Eq. (8.44) can be transformed to

$$\chi(x) = -\varkappa\int G(x, x')\varphi^2(x')dx'. \tag{8.45}$$

Substituting (8.40) and (8.45) to the second Eq. (8.44), we arrive, in the same slow motion approximation, at Eq. (8.42).

Let us note that the correction $\delta V = -(\alpha^2/2)\varphi^4$ to the potential has emerged from an analysis of the classical Eq. (8.44). On the other hand, quite the same answer can be obtained by calculating the first quantum correction to the potential of the φ field due to interaction with the χ field at zero 4-momenta of the external lines corresponding to a quanta of the φ field. The last term in the expression (8.42) is usually interpreted as "back reaction of radiation" [141].

During inflation, the field is regarded as homogeneous, $\varphi = \varphi(t)$, and Eq. (8.42) is greatly simplified. Let us write it for the specific case $u(\varphi) = \varphi^2$, also taking into account that the scale factor a is expressed in the usual way in terms of the Hubble parameter H, $a = \exp(\int H dt)$:

$$\frac{d^2\varphi}{dt^2} + 3H\frac{d\varphi}{dt} + V'(\varphi) + \frac{4\alpha^2}{m_\chi^2}\left[3H\varphi^2\frac{d\varphi}{dt} + \varphi^2\frac{d^2\varphi}{dt^2} + \varphi\left(\frac{d\varphi}{dt}\right)^2\right] = 0.$$

Due to a slow change of the φ field in time, we discard the terms proportional to $d^2\varphi/dt^2$ and $(d\varphi/dt)^2$, which is justified by the final result, and after that we arrive at the easily integrable equation

$$\left(3H + \frac{12H\alpha^2}{m_\chi^2}\varphi^2\right)\dot{\varphi} + V'(\varphi) = 0. \tag{8.46}$$

Solving this equation for the potential $V(\varphi) = \lambda\varphi^n$, taking into account the usual relationship between the Hubble parameter and the potential,

$$H = \sqrt{8\pi V(\varphi)/3}/M_{\rm Pl},$$

we arrive at an expression for the implicit time dependence of the field variable φ:

$$t = \frac{\sqrt{24\pi}}{M_{\rm Pl}\lambda^{1/2}}\left[\frac{1}{n(2-n/2)}\left(\varphi_0^{2-n/2} - \varphi^{2-n/2}\right) + \frac{4}{n(4-n/2)}\frac{\alpha^2}{m_\chi^2}\left(\varphi_0^{4-n/2} - \varphi^{4-n/2}\right)\right]. \tag{8.47}$$

Here the first term reproduces the result of the standard inflationary model while the second one takes into account the interaction with virtual χ-particles.

It is necessary to note that according to Eq. (8.46) there is an intermediate inflationary stage at which one already cannot neglect the second-order derivative in the φ field in the square brackets. Nevertheless, of greater interest is the first stage, and it is this stage that is studied below. The further consideration is performed under the assumption that the second term dominates in Eqs. (8.46) and (8.47), i.e., under the condition

$$2\alpha\varphi/m_\chi \gg 1. \tag{8.48}$$

For what follows, it is convenient to introduce the quantity

$$\varphi_c \equiv \frac{m_\chi}{2\alpha}, \tag{8.49}$$

which represents a boundary field value and separates the first stage of superslow motion of the φ field from the second stage of ordinary inflation. In both cases the field dynamics is described by simple analytical expressions.

In the case $\varphi > \varphi_c$, the time dependence of the field is

$$\begin{aligned} \varphi(t) &= \left[\varphi_0^{4-n/2} - t/f_n\right]^{1/(4-n/2)}, \\ f_n &= \frac{8\sqrt{6\pi}}{n(4-n/2)} \frac{\varkappa^2}{m_\chi^4 M_{\mathrm{Pl}} \lambda^{1/2}}. \end{aligned} \tag{8.50}$$

This expression is obtained under the "superslow-rolling" condition, which, according to Eq. (8.43), has a somewhat unusual form

$$\ddot{\varphi} \ll 12H\varphi^2\dot{\varphi}\frac{\alpha^2}{m_\chi^2}. \tag{8.51}$$

It is important to note that, as expected, the field changing rate

$$\dot{\varphi} = \frac{m^2}{12\alpha^2\varphi^2} \cdot \frac{V'}{H},$$

obtained from (8.46) with (8.48), turns out to be much smaller than the usual value $\dot{\varphi} = V'/3H$. The first stage of inflation terminates when the condition (8.51) does not hold any more. Further on follows a period of ordinary inflation which lasts as long as the condition $\ddot{\varphi} \ll 3H\dot{\varphi}$ holds. As estimates show, the second stage turns out to be very short.

Let us determine the values of quantum fluctuations emerging at the inflationary stage. The fluctuations of noninteracting fields have been studied well enough [272]. Meanwhile, for fields with self-interaction, only order-of-magnitude estimates are known [138].

Let us address the fluctuations emerging at the first stage of inflation for the potential $\lambda\varphi^4$. It can be done most simply if we introduce the auxiliary field $\widetilde{\varphi}$ and, by substituting $\widetilde{\varphi} = (\alpha/m_\chi)\varphi^2$, bring the action

(8.41) to the form

$$S = \int dx\sqrt{-g}\left[\partial_\mu\tilde{\varphi}\partial^\mu\tilde{\varphi} - \frac{1}{2}\tilde{m}^2\tilde{\varphi}^2\right], \tag{8.52}$$

corresponding to a free field with the mass $\tilde{m} \equiv m_\chi\sqrt{2\lambda}/\alpha$. Such a substitution is admissible at this stage of inflation where the field values are manifestly greater than zero. For a massive field, the fluctuation value is known [90], $\Delta\widetilde{\varphi} = \sqrt{3/(8\pi^2)}H^2/\tilde{m}$. Also known is the restriction on the mass of the field $\widetilde{\varphi}$, obtained by comparison with the measurements of energy density fluctuations by COBE [32], $\delta\rho/\rho \approx 6\cdot 10^{-5}$ on the modern horizon scale: $\tilde{m} \sim 10^{-6}M_{\rm Pl}$. Expressing $\tilde{m}$ in terms of the initial parameters, we obtain a relationship between them:

$$\frac{m_\chi}{M_{\rm Pl}}\frac{\sqrt{\lambda}}{\alpha} \sim 10^{-6}. \tag{8.53}$$

Since it is natural to suppose $m_\chi \ll M_{\rm Pl}$, the inequality (8.53), obtained from the observational data, is not burdensome and holds in a wide range of the parameters.

Let us determine the field value φ_U which gave rise to the causally connected region that created the visible part of the Universe. It is known that the number of e-folds necessary for explaining the observational data is $N_U \simeq 60$. Then, using, as always, the relationship $N_U = \int_{\varphi_U}^{\varphi_{\rm end}} H dt$, we get

$$\begin{aligned} N_U &= \int_{\varphi_U}^{\varphi_c}\frac{H(\varphi)}{\dot{\varphi}}d\varphi + \int_{\varphi_c}^{\varphi_{\rm end}}\frac{H(\varphi)}{\dot{\varphi}}d\varphi \\ &= \frac{2\pi\alpha^2}{M_{\rm Pl}^2 m_\chi^2}\left(\varphi_U^4 - \varphi_c^4\right) + \frac{\pi}{M_{\rm Pl}^2}(\varphi_c^2 - \varphi_{\rm end}^2). \end{aligned} \tag{8.54}$$

Taking into account the different time dependencies of the φ field at the first and the second stages of inflation, and also using the smallness of the field value $\varphi_{\rm end}$ right after the end of inflation as compared with its initial value φ_U, we obtain the desired expression

$$\varphi_U \simeq \left(\frac{N_U}{2\pi}\right)^{1/4}\sqrt{\frac{M_{\rm Pl}m_\chi}{\alpha}}. \tag{8.55}$$

We note that the visible part of the Universe in this case can form at $\varphi < M_{\rm Pl}$, at lower energies as compared to quadratic inflation. It is explained by the "superslow" field motion at the first stage of inflation, allowing the Universe to expand to a necessary size.

All the above reasoning is correct if the field χ has a sufficiently large mass, so that, at inflation, it can be considered as being located at the minimum of its potential. It is known that a field begins to roll rapidly to a minimum if the Hubble parameter becomes smaller than its mass, $H < m$. The Hubble parameter is time-dependent, therefore let us give the necessary estimates for the initial moment of emergence of the visible Universe ($\varphi = \varphi_U$) and at termination of the first stage of inflation ($\varphi = \varphi_c$). Simple calculations lead to the following result:

$$m_\chi > H(\varphi_U) \mapsto \frac{\sqrt{\lambda}}{\alpha} \le 0.1,$$
$$m_\chi > H(\varphi_c) \mapsto m_\chi \le M_{\text{Pl}} \frac{\sqrt{\lambda}}{\alpha^2}. \tag{8.56}$$

These inequalities hold in a wide range of the parameter, at least near the termination time of the first stage of inflation. Quantum corrections to the mass m_χ of the χ field bound the mass from below, therefore we will further assume that $m_\chi \ge \varkappa$, or, which is the same, $\alpha \le 1$.

Thus the interaction between the inflaton field and the quantum fluctuations of a massive field of another sort created by the inflaton makes it possible to explain, in particular, the smallness of energy density fluctuations under sufficiently soft conditions on the parameters of the potential.

8.5 Suppression of vacuum decay by virtual particles

There have probably been a few phase transitions in the early Universe. To such transitions one can attribute the one at the Grand Unification scale of $\sim 10^{17}$ GeV and the one at the electroweak interaction scale, ~ 100 GeV. So far we have been considering the class of models on the basis of chaotic inflation. However, there also exists an inflationary scenario on the basis of a first-order phase transition from which the Universe evolution begins [244]. In field theory, first-order phase transitions happen by means of a decay of a metastable (false) vacuum into spherically symmetric domains filled by a true vacuum. These domains begin to expand and occupy more and more volume. A calculation of the probability of such a decay represents a nontrivial problem.

This problem was repeatedly discussed in the literature in different areas of physics, such as field theory, solid state physics and the physics of phase transitions. Theoretical aspects of this phenomenon have been studied from a quantum-mechanical viewpoint for quite a long time [262]. Significant progress in these studies in quantum field theory began in the 70s, see [116], and these works laid the basis of the instanton approach to the calculations of vacuum decay. The same method is also applicable for calculating the probability of quantum-mechanical tunneling. The basic idea is that the integration contour is deformed in such a way that the basic contribution to the integral representing the transition amplitude be given by a single trajectory, the "instanton" $x_{\rm inst}(t_E)$, where t_E is time in Euclidean space. The subbarrier transition amplitude obtained in this way contains the necessary suppressing factor $\exp(-S[x_{\rm inst}])$.

Further progress in the subbarrier transition theory revealed the existence of numerous additional factors affecting the transition dynamics. These factors can appreciably change the physical picture of the transition and particular conclusions of one or other cosmological models.

One of the main factors is the temperature at which the subbarrier transition takes place. In cosmological phase transitions, the temperature corrections to the effective potential strongly distort its shape, so that at high temperatures such a transition can simply be absent [272]. As the temperature decreases, there emerge additional minima of the potential, and a phase transition becomes possible, though the parameters of the effective potential are different from those observed at present.

Another fundamental factor is the ambient medium. An interaction with it usually diminishes the tunnelling probability. When studying phase transitions due to false vacuum decay, it is necessary to take into account energy dissipation during the phase transition itself, at the motion of the wall that separates the true vacuum from the false one. It is of particular importance in studying the electroweak transition at the early stage of the Universe evolution [277], when the density of the ambient medium is high.

A third factor is interaction of a particle or a classical field with virtual particles of another sort, see [352]. The "heavy nucleon" model also confirms the conclusion that a cloud of scalar particles interacting with the nucleus affects the renormalizations of both its wave function and its mass.

A separate question is the calculation of quantum corrections to the vacuum decay probability in field theory. It turns out that the corrections, being small as compared with the main contribution, still lead to an appreciable change in the magnitude of the main effect, the subbarrier transition probability. The papers [252, 253] describe a method of summing the low levels which allows for a correct determination of the contribution of quantum corrections. In what follows, we consider the influence of a cloud of virtual particles on the classical evolution of a scalar field and the false vacuum decay probability.

We consider a first-order phase transition, where the scalar field making this transition is surrounded by a cloud of virtual particles of another sort. As will be shown below, such a cloud diminishes the probability of a subbarrier transition even at zero temperature, and the suppression can be strong enough to completely change the picture of the transition. It is sufficiently clear from a physical viewpoint because the cloud of virtual particles needs time for its state restructuring.

Consider a system with the action (8.37) in Euclidean space. Acting in full analogy with the previous sections, we obtain the effective Euclidean action for the main field φ:

$$S_{\text{eff}}^E = \int d^4x \left[-\frac{1}{2}\varphi_{,\mu}\varphi^{,\mu} - V(\varphi)\right] + \frac{\varkappa^2}{2}\int d^4x \int d^4x' u(\varphi(x)) G_E(x,x') u(\varphi(x')). \tag{8.57}$$

(Here and henceforth the gravitational effects are ignored). For what follows it is convenient to transform this expression to

$$S_E = \int d^4x \left[\frac{1}{2}(\partial\varphi)^2 + V(\varphi) - \frac{\alpha^2}{2}u(\varphi)^2\right] + \frac{\alpha^2}{2}\int d^4x \frac{\partial u(\varphi(x))}{\partial x_\mu}\frac{\partial u(\varphi(x))}{\partial x_\mu}. \tag{8.58}$$

The third term in this expression must be included in renormalization of the potential,

$$V_{\text{ren}}(\varphi) = V(\varphi) - \frac{\alpha^2}{2}u(\varphi(x))^2, \quad \alpha := \varkappa/m_\chi,$$

while the last term can substantially affect the system behaviour. It is this nonlocal term that reflects the influence of the cloud of virtual particles of the sort χ on the dynamics of the field ϕ. Here we also take into account that for the Green function $G_E(x-x')$ the inequality

$$\int dx G_E(x-x') = 1/m_\chi^2. \tag{8.59}$$

is valid.

Let the potential $V_{\rm ren}$ have two minima and the initial field value φ_F correspond to a metastable state. In this case, the metastable state decays by formation and expansion of true vacuum bubbles with the field value φ_T. This process is described by the $O(4)$-invariant solution $\varphi_B(r)$ of the classical field equation with the boundary conditions $\varphi_B(0) = \varphi_T, \varphi_B(\infty) = \varphi_F$. The decay probability density was found for the first time in the paper [116]:

$$\Gamma/V = \left(\frac{S_E(\varphi_B)}{2\pi}\right)^2 \left|\frac{\mathrm{Det}'\,\hat{D}(\varphi_B)}{\mathrm{Det}\,\hat{D}(\varphi_F)}\right|^{-1/2} \mathrm{e}^{-S_E(\varphi_B)}, \tag{8.60}$$

where the core K of the operator $\hat{D}(\varphi)$ is

$$K(x,y) \equiv \frac{\delta^2 S_E(\varphi)}{\delta\varphi(x)\,\delta\varphi(y)}. \tag{8.61}$$

The calculation of the normalizing determinant in the denominator is usually not difficult. Since $\varphi_F = \mathrm{const}$, everything is reduced to a calculation of the well-known determinant like $\mathrm{Det}(-\partial^2 + \mathrm{const})$. However, in the case under consideration, the effective action is nonlocal, and the calculation of the normalizing determinant is a problem. If we introduce the definitions

$$\Omega^2 \equiv V''(\varphi_F) + \alpha^2 \left(\frac{du}{d\varphi_F}\right)^2,$$
$$M^4 \equiv \varkappa^2 \left(\frac{du}{d\varphi_F}\right)^2, \tag{8.62}$$

then the problem reduces to calculating the determinant of the operator $\hat{D}(\varphi_F)$, whose core has the form

$$K_F(x,y) = \{\delta(x-y)(-\partial_x^2 + \Omega^2) - M^4 G_E(x-y)\}. \tag{8.63}$$

This core is not diagonal, but, due to its particular features, can be diagonalized and after that is easily calculated (see [259]). Let us introduce the operator $\hat{G}_E^{-1} \equiv (-\partial_x^2 + m_\chi^2)$. Then,

$$\begin{aligned} \mathrm{Det}\,\hat{D}(\varphi_F) &= \mathrm{Det}\,G_E^{-1}\hat{D}(\varphi_F)/\,\mathrm{Det}\,G_E^{-1} \\ &= \mathrm{Det}[(-\partial_x^2 + \Omega^2)(-\partial_x^2 + m_\chi^2) - M^4]/\,\mathrm{Det}\,G_E^{-1} \\ &= \mathrm{Det}(-\partial_x^2 + \omega_1^2)\cdot\mathrm{Det}(-\partial_x^2 + \omega_2^2)/\,\mathrm{Det}\,G_E^{-1}, \end{aligned} \tag{8.64}$$

where we have introduced the parameters

$$\omega_{1,2}^2 = \frac{1}{2}\left[\Omega^2 + m_\chi^2 - \sqrt{(\Omega^2 - m_\chi^2)^2 + 4M^4}\right]. \tag{8.65}$$

The problem of calculating the normalizing factor in the pre-exponential of the expression (8.60) has been reduced to calculations of known determinants of oscillator-type operators.

As a matter of fact, quantum corrections at high energies lead to serious alteration in the effective action and substantially affects the vacuum decay rate. Let us show that using as an example a potential with two minima, a local one and an absolute one,

$$V_{\mathrm{ren}} = \frac{\lambda}{8}(\varphi^2 - a^2)^2 + \frac{\varepsilon}{2a}(\varphi - a). \tag{8.66}$$

The instantonic solution to the Euclidean equation of motion can be characterized as follows:

$$\varphi(x) = \varphi_{\mathrm{inst}}(r) = A\tanh\left(\frac{M}{2}(r - R)\right) - B, \tag{8.67}$$

where $r^2 \equiv \sum_{\alpha=1}^4 x_\alpha^2$, and the parameters R and M are arbitrary quantities determined by minimization of the action. The parameters A and B are found from the boundary conditions

$$\varphi_{\mathrm{inst}}(r \to \infty) = \varphi_+, \quad \varphi_{\mathrm{inst}}(r \to 0) = \varphi_-. \tag{8.68}$$

Here φ_+ corresponds the right local minimum of the potential (8.66) (the false vacuum) and φ_- to the left, absolute minimum (the true vacuum).

The action (8.58) is written with neglected high-order derivatives in the field φ. This approximation is correct under the condition $\partial\varphi_B/m_\chi\varphi_B \ll 1$ (m_χ is the mass of χ particles that form the cloud). On the instanton trajectory, the derivative $\partial\varphi_B$ is proportional to the mass of a quantum of the main field, m_φ. Thus a validity criterion for the approximation is

$$m_\varphi/m_\chi \ll 1. \tag{8.69}$$

In fact, the transition domain from false to true vacuum is strongly widened by retarded effects due the cloud of virtual χ-particles. Therefore the above condition manifestly holds and is even too strict.

For making a numerical analysis, let us write down the classical equations whose solution is the O(4)-symmetric instantonic trajectory:

$$\partial_r^2\varphi + \frac{3}{r}\partial_r\varphi - V'_{\rm ren}(\varphi) + \frac{\alpha^2}{m_\chi^2}u'(\varphi)\left[\frac{3}{r}u'(\varphi)\partial_r\varphi + u''(\varphi)(\partial_r\varphi)^2 + u'(\varphi)\partial_r^2\varphi\right] = 0. \quad (8.70)$$

A solution to this equation is sought in the form (8.67) with the unknown parameters M and R. The parameters A and B are determined by the boundary conditions.

The results of the numerical calculations are shown in Fig. 8.1. The abscissa axis is chosen to coincide with the known result [116, 117], obtained without taking into account the cloud of virtual particles. It is seen from the plot that the action can increase by orders of magnitude and hence the phase transition probability turns out to be exponentially suppressed as compared with the usual predictions.

For a more specific estimate we assume that the action on the instantonic trajectory, without interaction with an additional field, is S_0, and that taking into account this interaction is ρS_0. Evidently, if this interaction is taken into account, the lifetime of the metastable state increases by

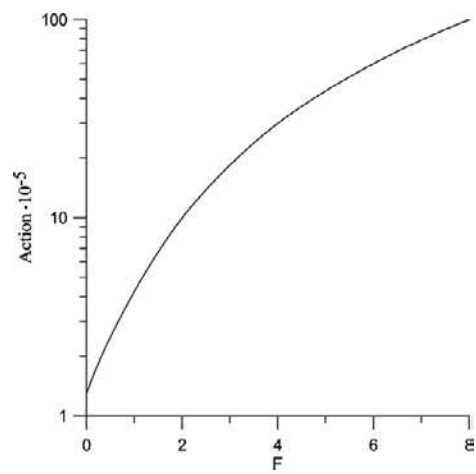


Figure 8.1 The dependence of the Euclidean effective action on the parameter $F = \varkappa^2/m_\chi^4$. The parameter values are $a = 1; \lambda = 0.1; \varepsilon = 0.01$ in units of $m_\chi = 1$.

a factor of $q \approx \exp[(\rho - 1)S_0]$. Using the plot of Fig. 8.1, we choose the intermediate value $\rho = 10$. The probability of a subbarrier transition is usually calculated in the quasiclassical approximation, such that $S_0 \gg 1$. Let us choose the minimally possible reasonable value of the action, $S_0 = 10$, which yields the estimate $q = \mathrm{e}^{90} \sim 10^{40}$. Such an abrupt increase in the lifetime of the metastable state is able to change the conclusions obtainable on the basis of a specific model. The action is usually much larger than the value chosen, and the suppression is still much stronger.

Evidently, the false vacuum decay probability is quite an uneasy problem, requiring the inclusion of first-order quantum corrections. The latter can drastically change the estimated lifetime of a metastable state and, in particular, affect the rate at which our Universe was formed if that happened due to a first-order phase transition, or the rate of the electroweak transition.

This page intentionally left blank

Chapter 9

The large-scale structure

In the first two sections of this chapter we briefly describe the modern views of galaxy formation. The subsequent sections are devoted to studying the new opportunities arising due to a correct consideration of quantum fluctuations during inflation.

9.1 The cosmic microwave background

In 1964, Arno Penzias and Robert Wilson decided to carry out some research in radioastronomy and satellite communication. For testing the antenna, they chose the wavelength of 7.35 cm. Very soon it became clear that the antenna permanently detected an unexplained additional noise of unclear origin, and no way was found to get rid of it. The isotropy of the noise and its constancy in time meant that its source was located somewhere outside the Solar system. Even if its origin were related to our Galaxy, its intensity would vary due to the Earth's rotation and orbital motion around the Sun, which change the antenna's direction in space. So it was clear that the noise was of extragalactic origin. On measurement, the intensity of this radio signal corresponded to the radiation intensity of an absolutely black body with a temperature near 3 K. That is how the Cosmic Microwave Background (CMB), also called the relic radiation, was discovered.

What is the CMB? According the Big Bang theory, the Universe emerged approximately 14 billion years ago as a result of an enormous "bang" that created our space and time and the whole existing matter and energy of our Universe. Up to an age of about 300 thousand years, the young Universe consisted of a hot plasma of elementary particles and, certainly, photons. The common expansion of the Universe gradually cooled this medium, and as soon as the temperature fell down to a value of a few thousand degrees, the free protons and electrons formed neutral atoms. After that, the existing radiation in the form of photons propagated freely since they almost do not interact with neutral atoms. Due to further cosmic expansion, the wavelengths of this radiation continually increased. It is this radiation that was detected by Pensias and Wilson as an unknown noise.

The CMB has a spectrum corresponding to the temperature $T = 2.73\,\mathrm{K}$. CMB investigations at different angles have revealed surprisingly small values of its temperature fluctuations, about 10^{-5} in relative units. Nevertheless, these fluctuations bear valuable information on the early Universe.

Let us discuss at a qualitative level the effects leading to the temperature fluctuations $\Delta T(\theta, \varphi)$. For convenience, the perturbations are expanded in spherical harmonics

$$\frac{\Delta T(\theta, \varphi)}{T} = \sum_{l,m} a_{lm} Y_{lm}(\theta, \varphi), \tag{9.1}$$

and the plot of the quantity $l(l+1)C_l$, where $C_l = \langle\, |a_{lm}|^2 \,\rangle$ is studied.

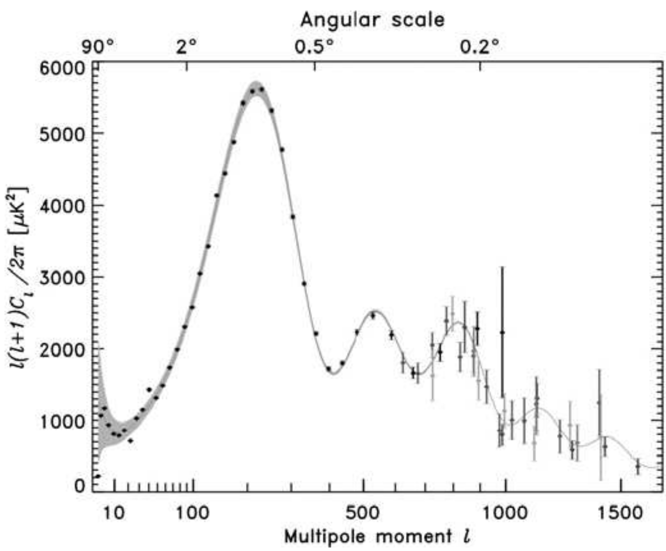


Figure 9.1 CMB temperature fluctuations.

This plot contains information of utmost importance. For example, we are now catching photons which were located near the horizon $l_{\rm rec}$ at recombination time, after which the Universe became transparent. (Recall that the horizon is the distance to which a light signal escapes in the expanding Universe from a certain time up to now.) While the Universe was opaque, the motion of massive particles was also hindered by friction due to interaction with photons.

On the other hand, in the same period of time there were fluctuations of the particle density and the gravitational potential. Subject to the gravitational field, the density fluctuations could have grown, but that did not happen because, as was said above, the motion of charged particles was strongly hindered. At recombination time, two events happened: (a) the friction disappeared, and particles began to move to the "centre" of the density fluctuation; in doing so, they are accelerated, which raises the local temperature; (b) photons from these heated regions move freely.

How does the size of the region of a density fluctuation affect further events?

1. Regions comparable with the horizon size at the recombination time, $L_{\rm rec} = 2H_{\rm rec}^{-1}$, are causally connected regions of the largest size at that time. These are regions having the largest gravitational well and hence the largest temperature contrast. Photons from such heated regions are now reaching us, and we must see a temperature peak in the plot. Let us estimate the size of these heated regions L_0 at present:

$$L_0 = \frac{1}{a(t)} L_{\rm rec} = \frac{T}{T_0} L_{\rm rec} \approx 300 \text{ Mpc}.$$

   Since $H_{\rm rec}^{-1} \approx \frac{3}{2} t_{\rm rec}$, we have $T_{\rm rec} \approx 3000\,\text{K}$.
   The angle at which we see such a region is

$$\theta = \frac{L_0}{L_U} \approx \frac{300}{9000} = 0.03 \text{ rad} \approx 2^\circ.$$

   Consequently, a peak must be observed at about 2°. This acoustic, or Doppler peak, is clearly seen in the plot.
2. Regions much larger than the horizon size $L_{\rm rec}$ at recombination evolve insignificantly. Their mean temperature weakly depends on the size, and there is an evident plateau in the plot (the Sachs–Wolfe plateau) at large angles.

3. Regions much smaller than L_{rec}. After recombination, these regions had enough time not only to contract but also to expand due to internal pressure. Therefore they are cooled a little, and we must see a minimum in the plot at smaller angles.

Classical evolution of quantum fluctuations

One of the most important results of the previous section is the following. At the inflationary stage, the density fluctuations created by quantum fluctuations rapidly increase their spatial size. At some instant, they become larger than the horizon size — the fluctuation crosses the horizon for the first time. By the end of inflation, the fluctuation scale can greatly exceed the horizon. The fluctuation magnitude is changing until the fluctuation scale crosses the horizon for the second time, now after the end of inflation. It turns out that in the period between two horizon crossings, there holds a simple relationship

$$\frac{\delta\rho}{p+\rho} \simeq \text{const}, \tag{9.2}$$

which we shall need in what follows. Let us first of all prove this relationship using Eq. (7.10) written in the form

$$\frac{d\rho}{p+\rho} = -\frac{3da}{a}.$$

Suppose there is a one-to-one correspondence between matter density and pressure, $\rho \leftrightarrow p$. Hence the pressure is a certain function of the density, $p(\rho)$, and the previous expression can be formally integrated:

$$\int_{\rho(t_0)}^{\rho(t)} \frac{d\rho}{p(\rho)+\rho} = -3\int_{a(t_0)}^{a(t)} \frac{da}{a}, \tag{9.3}$$

where t_0 is a time instant at the inflationary stage, and t is an arbitrary time satisfying the condition $t - t_0 \gg 1/H_e$, and H_e is the Hubble parameter at the end of inflation. If a quantum fluctuation creates a perturbation of the energy density $\delta\rho(t_0)$, then Eq. (9.3) should read

$$\int_{\rho(t_0)+\delta\rho(t_0)}^{\rho(t)+\delta\rho(t)} \frac{d\rho}{p(\rho)+\rho} = -3\int_{a(t_0)+\delta a(t_0)}^{a(t)+\delta a(t)} \frac{da}{a}, \tag{9.4}$$

which holds for the spatial region occupied by the fluctuation.

Let t_0 and t denote the times of the first and second horizon crossings by the fluctuation in question. Let us split both parts of Eq. (9.4) into a sum of integrals:

$$\int_{\rho(t_0)+\delta\rho(t_0)}^{\rho(t_0)} \frac{d\rho}{p(\rho)+\rho} + \int_{\rho(t)}^{\rho(t)+\delta\rho(t)} \frac{d\rho}{p(\rho)+\rho}$$
$$= -3\int_{a(t_0)+\delta a(t_0)}^{a(t_0)} \frac{da}{a} - 3\int_{a(t)}^{a(t)+\delta a(t)} \frac{da}{a}.$$

Here, Eq. (9.3), which is valid for a volume including the volume occupied by the fluctuation, is used in order to cancel the main terms in both parts of the equality. The remaining integrals can be easily estimated taking into account the smallness of the fluctuations:

$$\left(\frac{\delta\rho}{p+\rho} + 3\frac{da}{a}\right)_{t_0} \simeq \left(\frac{\delta\rho}{p+\rho} + 3\frac{da}{a}\right)_t .$$

The scale factor a grows exponentially since the beginning of the inflationary stage, so that we can neglect the second terms in both parts of the equation to obtain the required equality (9.2).

Equation (9.2), rewritten in the form

$$\left(\frac{\delta\rho}{p+\rho}\right)_{t_\mathrm{f}} \simeq \left(\frac{\delta\rho}{p+\rho}\right)_{t_\mathrm{in}}, \tag{9.5}$$

can be used for obtaining the spectrum of large-scale fluctuations. Here, $t_\mathrm{in} = t_0$ is the fluctuation emergence time at the inflationary stage, before the first horizon crossing, and t_f is the second horizon crossing time after the end of inflation at the matter-dominated stage. The left-hand side of this equality can be expressed in terms of the scalar field (inflaton) $\varphi = \varphi(t_\mathrm{in})$. The right-hand side of the equality is especially simple at the matter-dominated stage because the pressure is zero, $p = 0$. Let us write down the expression for $\delta\rho$ and $p+\rho$ at the inflationary stage:

$$\delta\rho = V'(\varphi)\delta\varphi, \qquad p+\rho = \dot{\varphi}^2.$$

Now, the quantity $\delta\rho/\rho|_{t_\mathrm{f}}$ to be estimated is also expressed in terms of the inflaton field:

$$\left(\frac{\delta\rho}{\rho}\right)_{t_\mathrm{f}} \simeq \left(\frac{V'(\varphi)\delta\varphi}{\dot{\varphi}^2}\right)_{t_\mathrm{in}}.$$

In the slow-rolling approximation we have $\delta\varphi \approx H(\varphi)/2\pi$. Using the inflaton field equation, we obtain

$$\left(\frac{\delta\rho}{\rho}\right)_{t_\mathrm{f}} = \left(\frac{9H(\varphi)^3}{2\pi V'(\varphi)}\right)_{t_\mathrm{in}}. \tag{9.6}$$

At the inflationary stage, the Hubble parameter H is expressed in terms of the potential $V(\varphi)$, and we arrive at a formula for calculating the density fluctuation magnitude:

$$\left(\frac{\delta\rho}{\rho}\right)_{t_\mathrm{f}} = \frac{9}{5}\left(\frac{4}{3}\right)^{3/2}\frac{m\varphi^2}{M_\mathrm{Pl}^3}. \tag{9.7}$$

The fluctuations are characterized by values of the φ field at first horizon crossing. This formula includes, for completeness, the factor $2/5$, which emerges when taking into account matter domination at the second horizon crossing. For what follows this factor is not very significant.

Another important parameter of the fluctuation is its characteristic size l. It depends on the times t_in and t_f of the first and the second horizon crossings, respectively. The fluctuation size at the time t such that $t_\mathrm{in} < t < t_\mathrm{f}$ can be estimated using a normalization to the size of the visible part of the Universe, L_U:

$$l = L_U \exp(N - N_U), \tag{9.8}$$

where N is the number of e-folds at which the scale l of this fluctuation is formed at the inflationary stage, and L_U denotes the size of our Universe, $L_U \approx 10^4\mathrm{Mpc} \approx 10^{28}$ cm.

On the other hand, the number of e-folds is expressed in terms of the inflaton value φ according to the expression (8.5):

$$N \approx \frac{2\pi}{M_\mathrm{Pl}^2}\varphi^2,$$

where the potential is chosen in the form $V(\varphi) = (m/2)\varphi^2$, and the inequality $\varphi \equiv \varphi_\mathrm{in} \gg \varphi_\mathrm{f}$ is assumed. Substituting this expression and Eq. (9.8) into the expression (9.6), we arrive at a relationship between the fluctuation magnitude $(\delta\rho/\rho)_{t_\mathrm{f}}$ and its size l [264]:

$$\begin{aligned}\left(\frac{\delta\rho}{\rho}\right)_{t_\mathrm{f}} &= \left(\frac{\delta\rho}{\rho}\right)_{t_U}\left[1 + \frac{1}{N_U}\ln\left(\frac{l}{L_U}\right)\right] \\ &\simeq \left(\frac{\delta\rho}{\rho}\right)_{t_U}\left(1 + \ln\frac{l}{L_U}\right)^{1/N_U} \simeq \left(\frac{\delta\rho}{\rho}\right)_{t_U}\left(\frac{l}{L_U}\right)^{1/N_U}.\end{aligned} \tag{9.9}$$

Here it is taken into consideration that the second term in the brackets is much smaller than unity at large scales, such that $l/L_U \geq 0.01$.

Consequently, the magnitudes of large-scale fluctuations very weakly depend on their spatial size. In this case, one speaks of an almost flat fluctuation spectrum. The above series expansion in $l \approx L_U$ leads to a power-law spectrum with the spectral index $n \equiv 1 - 2/N_U < 1$. Let us note that one can judge on the perspective of a specific model of inflation by the observed value of the spectral index.

9.2 The development of density fluctuations

9.2.1 Density fluctuations in Minkowski space

Let us follow the development of energy density fluctuations. We will suppose that the substance which is the carrier of energy can be considered as a perfect fluid with the stress-energy tensor

$$T_{\mu\nu} = pg_{\mu\nu} + (p+\rho)U_\mu U_\nu.$$

Here U_μ are the velocity components of an elementary volume of the fluid. In the comoving reference frame,

$$U^t = 1, \quad U^i = 0.$$

The hydrodynamic equations of motion have the form

$$\frac{\partial \rho}{\partial t} + \nabla(\rho \bar{v}) = 0,$$
$$\frac{\partial \bar{v}}{\partial t} + (\bar{v}\nabla)\bar{v} + \frac{1}{\rho}\nabla p + \nabla\varphi = 0,$$
$$\Delta\varphi = 4\pi G\rho, \tag{9.10}$$

where φ is the Newtonian gravitational potential. The solution of this equation in the zero-order approximation is elementary:

$$\bar{v}_0 = 0, \quad \rho_0 = \text{const}, \quad p_0 = \text{const}.$$

In general, it is just an approximate solution, satisfactory when the influence of the gravitational field potential is small. Consider the behaviour of small fluctuations, marked with the subscript 1. Equations (9.10) are simplified:

$$\frac{\partial \rho_1}{\partial t} + \nabla(\rho_0 \bar{v}_1) = 0,$$
$$\frac{\partial \bar{v}_1}{\partial t} + \frac{v_s^2}{\rho_0}\nabla\rho_1 + \nabla\varphi_1 = 0,$$
$$\Delta\varphi_1 = 4\pi G\rho_1. \tag{9.11}$$

The velocity of sound is $v_s^2 \equiv (\partial p/\partial\rho)_{\text{adiabat}} \simeq p_1/\rho_1$.

From Eqs. (9.11) it is easy to obtain the wave equation

$$\frac{\partial^2 \rho_1}{\partial t^2} - v_s^2 \nabla \rho_1 = 4\pi G \rho_0 \rho_1,$$

whose stability must be investigated. To this end we seek a solution in the form of plane waves

$$\rho_1(\bar{r}, t) = A\rho_0 \, \mathrm{e}^{-i\bar{k}\bar{r} + i\omega t}.$$

From the expression for the frequency

$$\omega = \pm\sqrt{k^2 v_s^2 - 4\pi G \rho_0}$$

it is evident that at small frequencies, such that

$$k < k_J \equiv \sqrt{4\pi G \rho_0 / v_s^2},$$

the solution is unstable.

The theory of gravitational instability was discussed for the first time by James Jeans in his paper "Stability of a spherical nebula" (1902), which considered quantitatively the gravitational instability of a gas cloud. The index "J" appearing in this section is related to Jeans's name.

The growing modes destroy the zero-order approximation in the form of a homogeneous energy density distribution. Under the impact of gravity, the structures become more dense. There occurs an exponential growth of fluctuations, ending with the formation of stars and stellar clusters. The size of an unstable region is, by order of magnitude,

$$l \equiv l_J > 2\pi / k_J.$$

One can also estimate the Jeans mass of matter M_J, above which the structure turns out to be unstable:

$$M_J \equiv \frac{4\pi}{3} \left(\frac{\pi}{k_J} \right)^3 \rho_0 = \frac{\pi^{5/2}}{6} \frac{v_s^2}{G^{3/2} \rho_0^{1/2}}.$$

9.2.2 Density perturbations in the expanding Universe

In the previous section, the cosmological expansion was not taken into account. Therefore the conclusions are correct only in a limited number of situations, for instance, for intragalactic gas condensation.

In a homogeneous Universe, the expansion is described by the time dependence of the scale factor $a(t)$. Then Eqs. (9.10) preserve their form, but their solution changes drastically. As a zero-order approximation, we will choose a solution in the form

$$\rho_0(t) = a^{-3}(t)\rho_0(t_0), \quad \vec{v}_0(t) = H(t)\vec{r}, \quad \nabla\varphi_0 = \frac{4\pi G}{3}\rho_0(t)\vec{r}.$$

We introduce small perturbations marking them by the subscript 1. Acting as above, we obtain

$$\ddot{\delta}_k + 2H\dot{\delta}_k + \left(\frac{v_s^2 k^2}{a^2(t)} - 4\pi G\rho_0(t)\right)\delta_k = 0,$$

where δ_k is the Fourier image of the relative density $\delta\rho/\rho \simeq \rho_1/\rho_0$. The velocity of sound v_s is proportional to the velocity of particles, therefore $v_s \sim 1/a(t)$. An unstable regime emerges at large wavelengths, i.e., at

$$k < k_J \equiv a(t)\sqrt{4\pi G\rho_0(t)/v_s^2(t)}.$$

Consider the growth of long-wave fluctuations, such that $k \ll k_J$. Recalling that at the matter-dominated stage we have the relationships

$$H(t) = \frac{\dot{a}}{a} = \frac{2}{3t}, \quad \rho_0(t) = \frac{1}{6\pi G t^2},$$

we obtain from the original equation that

$$\ddot{\delta}_k + \frac{4}{3t}\dot{\delta}_k - \frac{2}{3t}\delta_k = 0,$$

which leads to growing solutions:

$$\delta_k(t) = \delta_k(t_{\rm in})\left(\frac{t}{t_{\rm in}}\right)^{2/3}.$$

The long-wave fluctuations, having been formed, grow much slower than in a stationary Universe. Further on, these density fluctuations lead to the formation of galaxies and clusters of galaxies.

9.3 The baryonic asymmetry of the Universe

It is well known that, despite the symmetry between particles and antiparticles, baryons dominate in our Universe. It is one of the most important ingredients of the modern cosmological picture. At which moment did such an essential symmetry breakdown occur? What is the mechanism of baryosynthesis? A great number of reviews are devoted to this problem, see, e.g., [138, 232]), where quite various opportunities are discussed.

The whole previous presentation assumed homogeneity (on the average) of energy density and entropy fluctuations. In this chapter we will show that, under certain conditions, large-scale fluctuations of different kinds are also possible. This leads to new phenomena, such as macroscopically large antimatter regions and massive primordial black holes. The conditions for such phenomena originate in the course of inflation and develop in the post-inflationary period. The possible formation of antimatter regions has also been discussed earlier [110].

Since antimatter islands (regions) have not yet been found experimentally, the number of antibaryons in the whole Universe must be much smaller than the number of baryons. For the same reason, antimatter islands cannot be too large — otherwise we would observe annihilation photons at the edges of the corresponding regions.

On the other hand, if the size of antimatter regions is not sufficiently large, then annihilation at their boundaries should lead to their rapid evaporation, and it is just what happens in the majority of models of baryosynthesis . The processes of spontaneous CP symmetry breaking usually manifest themselves in the course of first- and second-order phase transitions that took place in the early Universe. In this case, the antimatter regions turn out to be too small to be preserved until the present [138]. As has been shown in [233], a boundary separating an antimatter region from its baryonic environment moves due to annihilation not farther than by 0.5 pc by the termination of the radiation-dominated epoch. Thus, to be preserved up to the present time, the antimatter regions should be large enough. This in turn means that the future antimatter regions should have begun their formation at the inflationary stage, because the main property of this stage is the ability to effectively “stretch” the spatial lengths.

9.3.1 Baryogenesis

From the large number of models of baryosynthesis, we would like to choose the one [138, 139] where the inflationary period plays a substantial role.

This model postulates the existence of a complex scalar field χ, possessing a baryonic charge. Since at present the baryonic charge is concentrated in fermions (quarks), it is necessary to provide a transition of the baryonic charge from the χ field to the quarks. To this end, one introduces the interaction Lagrangian of the χ field with heavy quarks Q and leptons L, so that the full Lagrangian has the form

$$L = -\partial_\mu \chi^* \partial^\mu \chi - V(\chi) + i\bar{Q}\gamma^\mu \partial_\mu Q + i\bar{L}\gamma^\mu \partial_\mu L \\ - m_Q \bar{Q}Q - m_L \bar{L}L + (g\chi\bar{Q}L + \text{h.c.}). \tag{9.12}$$

It is supposed that the heavy quarks Q and leptons L are connected with the ordinary quarks and leptons of matter fields. The fields χ and Q possess a baryon number but not a lepton number. On the contrary, the field L possesses the lepton number only. Consequently, the term in the Lagrangian (9.12), determining their interaction, leads to lepton number nonconservation [139]. The baryon number conservation follows from the $U(1)$ symmetry of the Lagrangian with respect to the transformations

$$\chi \to \exp(i\beta)\chi, \quad Q \to \exp(i\beta)Q, \quad L \to L. \tag{9.13}$$

Let us return to studying the dynamics of the scalar field χ that creates the baryon asymmetry. We choose the potential in the first-order approximation in the form

$$V(\chi) = -m_\chi^2 \chi^* \chi + \lambda_\chi (\chi^* \chi)^2 + V_0 = \lambda(|\chi|^2 - f/\sqrt{2})^2 + V_0, \tag{9.14}$$

where the constant V_0 is added for the minimum of the potential to be equal to zero. The "Mexican hat" potential (9.14) is the first, basic approximation. In what follows we will need quantum corrections to it.

The χ field can be presented in the form

$$\chi(\vartheta) = \frac{f}{\sqrt{2}} \exp\left(\frac{i\vartheta}{f}\right). \tag{9.15}$$

A violation of the $U(1)$ symmetry means that the radial component of the χ field takes the fixed value

$$f = m_\chi / \sqrt{\lambda_\chi},$$

and the variable ϑ in Eq. (9.15) now has the meaning of a massless scalar field since $V(\vartheta) = \text{const}$. The minimum of the potential is located on a circle of radius f. Further on we work with the dimensionless variable $\theta = \vartheta/f$.

The quantum corrections change the shape of the potential, so that its symmetry (9.13) is violated, and a small tilt of the "Mexican hat" is observed:

$$V(\theta) = \Lambda^4(1 - \cos\theta); \quad \Lambda \ll f. \tag{9.16}$$

This potential has a countable number of minima at $\theta = 2\pi N$, $N = 0, 1, 2, \ldots$. Note that the potential (9.16) is a convenient approximation of a more involved expression, see, e.g., [4]. The final form of our potential is

$$V_b(\chi) = \lambda(|\chi|^2 - f/\sqrt{2})^2 + \Lambda^4(1 - \cos\theta) + V_0. \tag{9.17}$$

For baryogenesis, of importance is the tilt of the potential (9.16) near the minimum which fixes the mass of the θ field:

$$m_\theta^2 = \frac{\Lambda^4}{f^2}. \tag{9.18}$$

Since the tilt is caused by quantum corrections, it is natural to assume a small value of the parameter Λ. It means that the contribution of the field θ to the energy density is small as compared with the full energy density at the inflationary stage. For this reason, the field θ is effectively massless, so that

$$m_\theta \ll H, \tag{9.19}$$

where H is the Hubble parameter at inflation. After the end of inflation, when the condition (9.19) is violated, the field θ oscillates near the minimum of the potential (9.16), and the energy $\rho_\theta \simeq \theta_i^2 m_\theta^2 f^2$ of the field θ passes on to the energy of created baryons and antibaryons [114, 139]. The sign of the baryonic charge in the final state depends on the initial value of phase at the time of symmetry breaking.

As a result of the symmetry breaking, the phase β of the χ field acquires the random value θ, and the effective Lagrangian has the form

$$\begin{aligned} L = &-\frac{f^2}{2}\partial_\mu\theta\partial^\mu\theta + i\bar{Q}\gamma^\mu\partial_\mu Q + i\bar{L}\gamma^\mu\partial_\mu L \\ &- m_Q\bar{Q}Q - m_L\bar{L}L + \left(\frac{g}{\sqrt{2}}f\bar{Q}L + \text{h.c.}\right) + \partial_\mu\theta\bar{Q}\gamma^\mu Q. \end{aligned} \tag{9.20}$$

For a spatially homogeneous field $\chi = (f/\sqrt{2})\,\mathrm{e}^{i\theta}$, we have the following expression for the formed baryonic charge:

$$Q = i[\chi^* d\chi/dt - (d\chi^*/dt)\chi] = -f d\theta/dt. \tag{9.21}$$

It is clear that $Q > 0$ if $\dot{\theta} < 0$ in the process of classical motion to the zero of the phase θ. Consequently, a clockwise motion leads to a baryon excess, and a counterclockwise motion to an antibaryon excess.

Let us estimate the number of baryons and antibaryons created at oscillations of the phase θ with an arbitrary initial phase θ_i. An expression for the concentration of the baryons formed, $n_{B(\overline{B})}$, in the limit of a small phase θ_i has the form [139]

$$n(Q, L) = \frac{1}{V}\sum_{s_Q, s_L}\int \frac{d^3p}{(2\pi)^3 2p^0}\frac{d^3q}{(2\pi)^3 2q^0}|\langle Q(p, s_Q)L(q, s_L)|0\rangle|^2. \tag{9.22}$$

After standard calculations using the last term in (9.12) as an interaction Lagrangian, we obtain the density of created baryons

$$n_{B(\overline{B})} = \frac{g^2}{\pi^2} \int_{m_Q+m_L}^{\infty} \omega d\omega \left| \int_{-\infty}^{\infty} dt \chi(t)\, \mathrm{e}^{\pm 2i\omega t} \right|^2 , \tag{9.23}$$

which is correct in the case $\chi(t \to -\infty) = \chi(t \to +\infty) = 0$.

For the more general case $\chi(t \to -\infty) \neq 0$, $\chi(t \to +\infty) = 0$ a result can be obtained by partial integration of the expression (9.23), which leads to

$$N_{B(\bar{B})} = \frac{g^2}{4\pi^2} \Omega_{\theta_i} \int d\omega \left| \int_{-\infty}^{\infty} d\tau \dot{\chi}(\tau) e^{\pm 2i\omega\tau} \right|^2 , \tag{9.24}$$

where Ω_{θ_i} is a volume containing the initial phase θ_i. In obtaining Eq. (9.24), the surface term is equal to zero at $t = \infty$ due to Feynman conditions on radiation of the χ field at $t = -\infty$.

A numerical result can be approximately found by assuming the time dependence of the phase in the form

$$\theta(t) \approx \theta_i (1 - m_\theta t), \tag{9.25}$$

during the first period of oscillations. Substituting (9.25) and (9.15) into (9.24), we obtain after simple calculations:

$$N_{B(\bar{B})} \approx \frac{g^2 f^2 m_\theta}{8\pi^2} \Omega_{\theta_i} k(\theta_i), \quad k(\theta_i) = \theta_i^2 \int_{\mp \frac{\theta_i}{2}}^{\infty} d\omega \frac{\sin^2 \omega}{\omega^2}, \tag{9.26}$$

where the sign at the lower limit of the integral corresponds to a baryonic or antibaryonic excess.

At the reheating time, the inflaton energy density is transformed to the energy density of relativistic particles. It is then assumed that the inflaton decays into light particles rather rapidly as compared with the oscillation period near a minimum (recall that due to the tilt of the potential the minimum is located at the point $\theta = 0$). Thus we have $\Gamma_{\mathrm{tot}} \gg m_\theta$. Oscillations of the field θ begin when friction in the medium can be neglected, i.e., at $H \approx m_\theta$. The phase change with time leads to baryon and antibaryon creation as discussed above. The entropy density after thermalization is

$$s = \frac{2\pi^2}{45} g_* T^3, \tag{9.27}$$

where g_* is the total effective number of degrees of freedom. It is assumed that the temperature exceeds the electroweak symmetry violation scale.

At this temperature all degrees of freedom of the Standard Model are in equilibrium and $g_* \simeq 106.75$. The temperature is connected with the expansion rate as follows:

$$T = \sqrt{\frac{M_{\rm Pl} H}{1.66 g_*^{1/2}}} \approx \frac{\sqrt{M_{\rm Pl} m_\theta}}{g_*^{1/4}}. \tag{9.28}$$

In the last expression (9.28) it is assumed that relaxation processes begin when the condition $H \approx m_\theta$ holds. Using Eqs. (9.26), (9.27) and (9.28), we obtain different expressions for baryon and antibaryon concentrations:

$$\frac{n_{B(\bar{B})}}{s} = \frac{45 g^2}{16\pi^4 g_*^{1/4}} \left(\frac{f}{M_{\rm Pl}}\right)^{3/2} \frac{f}{\Lambda} k(\theta_i). \tag{9.29}$$

The function $k(\theta_i)$ includes the dependence of the magnitude and sign of the baryon asymmetry on the initial phase which is different in different spatial regions. The behaviour of this function can be easily analyzed numerically with the aid of the expression (9.26).

The expression (9.29) makes it possible to obtain the observed average baryonic asymmetry in the Universe, $n_B/s \approx 3 \cdot 10^{-10}$. In this model it has been assumed that $f \geq H \simeq 10^{-6} M_{\rm Pl}$. A natural value of the coupling constant is $g \leq 10^{-2}$, which leads to the observable baryon asymmetry under the reasonable requirement $f/\Lambda \geq 10^5$ [4].

9.3.2 Large-scale fluctuations of the baryonic charge

There appears an opportunity of interest if one takes into account the effect of phase fluctuations during inflation. To do so, let us consider in more detail the motion of the phase along the valley $|\chi| = f/\sqrt{2}$, having the shape of a circle. Let the phase $\theta = 0$ correspond to the north pole and $\theta = \pi$ to the south pole. The minimum of the potential is located at the north pole. Using Eq. (9.21), it is easy to show that preferred antibaryon production takes place if the field is moving to the north pole *counterclockwise*, while baryon production occurs if the field is moving *clockwise*.

Let us calculate the volume distribution of antimatter domains. The total volume of all antimatter domains formed by the end of the N_t-th e-folding before the end of inflation may be calculated using the recursive procedure: suppose the total volume of all the domains with the average phase $\bar{\theta}$ formed by that time is $V(\bar{\theta}, N_t)$. Then their total volume at the

$(N_t - 1)$-th e-folding before the end of inflation is given by

$$V(\bar{\theta}, N_t - 1) = e^3 V(\bar{\theta}, N_t) + [V_U(N_t) - e^3 V(\bar{\theta}, N_t)] \cdot P(\bar{\theta}, N_t - 1) \cdot h, \tag{9.30}$$

where $V_U(N_t)$ is the volume of the Universe $V_U(N_t) \approx \mathrm{e}^{3\cdot(N_U - N_t)} H^{-3}$ at N_t e-folds before the end of inflation, $N_U \approx 60$ is the total number of e-folds during inflation, and $h = H/(2\pi f)$. $P(\bar{\theta}, N_t - 1)$ gives the phase distribution, which is Gaussian, see (8.33), (8.30):

$$P(\bar{\theta}, N_t) = \frac{1}{\sqrt{2\pi}\sigma_{N_t}} \cdot \exp\left(-\frac{(\theta_U - \bar{\theta})^2}{2\sigma_{N_t}^2}\right), \tag{9.31}$$

$$\sigma_{N_t} = \frac{H}{2\pi \cdot \overline{|\chi(N_t)|}} \cdot \sqrt{N_U - N_t}. \tag{9.32}$$

Here θ_U is the initial phase of the field.

Note that here N_t is the number of e-folds ***before the end of inflation***, so it ***decreases*** with time; so if the moment $(t = 0)$ corresponds to the beginning of inflation, then $N_0 \equiv N_U \approx 60$, while at the end of inflation $N_\tau = 0$. Accordingly, the volume of antimatter domains at the beginning of inflation is $V(\bar{\theta}, N_U) = 0$.

The first term in Eq. (9.30), $\mathrm{e}^3 V(\bar{\theta}, N_t)$, is the total volume of antimatter domains formed ***before*** the N_t-th e-folding. The second term

$$v(\bar{\theta}, N_t) = [V_U(N_t) - e^3 V(\bar{\theta}, N_t)] \cdot P(\bar{\theta}, N_t - 1) \cdot h, \tag{9.33}$$

is the total volume of antimatter domains formed ***during*** the N_t-th e-folding. Since the initial volume of each domain is H^{-3}, the number of domains formed during a given e-folding is

$$n = \frac{v(\bar{\theta}, N_t)}{H^{-3}}. \tag{9.34}$$

Domains grow in size during inflation, so the earliest domains to form become the biggest at the end of inflation. Their linear sizes at present are determined by the equation

$$L(N_t) = 6 \cdot 10^3 e^{-(N_U - N_t)}. \tag{9.35}$$

Here $L(N_t)$ is the size (in Mpc) of the antimatter domain formed at the N_t-th e-folding.

Using the relations obtained, one can calculate the size distribution of matter and antimatter domains. However, to be realistic, our model should suppress in a consistent way the large-scale fluctuations of the baryonic charge. Otherwise such fluctuation would contribute to the CMB spectrum and would be observed. Indeed, according to the mechanism we are discussing, the CMB temperature fluctuations are proportional to the energy density of matter, $\delta\rho/\rho \simeq \delta\theta\Omega_B/\Omega_{\rm tot}$. Here $\delta\theta = H/(2\pi f)$, see above, Ω_B is the relative baryonic energy density, and $\Omega_{\rm tot}$ is the total relative energy density. Consequently, we have the restriction $\delta\theta \leq 10^{-3}$, at least at large scales. But in this case, for 60 e-folds during inflation, the phase will move by an angle of at most $\Delta\theta \sim 60 \cdot 10^{-3} \sim 0.1$ rad to zero, and a transition through the zero value is very unlikely.

Let us discuss how this difficulty could be overcome.

The first method is to take into account the interaction of χ with the inflaton φ, with a certain coupling constant g [234]. Then the potential (9.14) takes the form

$$V(\varphi,\chi) = \lambda(|\chi|^2 - f^2/2)^2 - g|\chi|^2(\varphi - cM_{\rm Pl})^2, \qquad (9.36)$$

where λ, g and c are parameters of the potential. This potential also has the shape of a Mexican hat with a maximum at

$$|\chi| \equiv f_{\rm eff}(\varphi) = \sqrt{f^2 + \frac{g}{\lambda}(\varphi - cM_{\rm Pl})^2}. \qquad (9.37)$$

Now, evidently, the position of the minimum is not constant. On the contrary, it strongly depends on the classical field φ, which changes in the course of inflation,

$$\varphi(t) = \varphi_U - \frac{m_\varphi M_{\rm Pl}}{2\sqrt{3\pi}}t, \qquad (9.38)$$

for the quadratic potential $U(\varphi) = m_\varphi^2\varphi^2/2$. Recall that the fluctuation magnitude of the phase θ of the field χ is inversely proportional to the quantity $f_{\rm eff}(\varphi)$. It is not hard to obtain the dependence of the inflaton value on the number of e-folds N:

$$\varphi = \varphi_N = \varphi_U - \frac{M_{\rm Pl}}{2\sqrt{3\pi}}N. \qquad (9.39)$$

Here it is taken into account that $m_\varphi \cong H$ and $N = Ht$. Thus in N e-folds after Universe formation the effective scale $f_{\rm eff}$ (9.37) looks as follows:

$$f_{\rm eff}(N) = f\sqrt{1 + \frac{g}{\lambda}\frac{M_{\rm Pl}^2}{f^2}\left[\left(\frac{\varphi_U}{M_{pl}} - c\right) - \frac{N}{2\sqrt{3\pi}}\right]^2}. \qquad (9.40)$$

Denoting $(\varphi_U/M_{pl} - c) \equiv N_f/(2\sqrt{3\pi})$, we arrive at the final expression of the form

$$f_{\text{eff}}(N) = f\sqrt{1 + \frac{g}{12\pi\lambda}\frac{M_{\text{Pl}}^2}{f^2}(N_f - N)^2}. \tag{9.41}$$

This expression has an important property: there is from the very beginning the large parameter $M_{\text{Pl}}^2/f^2 \sim 10^{10}$, and the function $f_{\text{eff}}(N)$ has a sharp minimum at $f_{\text{eff}}(N_f) = f$ at a reasonable relationship between the parameters. Hence it immediately follows that the fluctuation magnitude of the phase θ

$$\langle\,\delta\theta\,\rangle = \frac{H}{2\pi f_{\text{eff}}(N)}, \tag{9.42}$$

abruptly increases near the e-folding $N = N_f$.

Let us estimate the possible range of the parameters g and λ. The first inequality follows from the requirement that fluctuations should be sufficiently strongly suppressed, i.e., $f_{\text{eff}} \gg f$, and from the expressions (9.37):

$$\langle\,\chi\,\rangle \sim \sqrt{g/\lambda}M_{\text{Pl}} \gg f. \tag{9.43}$$

The second inequality is connected with the fact that the additional term inserted into the potential also renormalizes the inflaton mass,

$$m_{\text{ren}}^2 = m^2 - g\langle\,\chi^2\,\rangle.$$

The requirement that the renormalization should be small leads to the inequality $m^2 \gg g\langle\,\chi^2\,\rangle \sim g^2 M_{\text{Pl}}^2/\lambda$. Thus the second inequality reads

$$m \gg gM_{\text{Pl}}/\sqrt{\lambda}. \tag{9.44}$$

Choosing the numerical values of the parameters

$$f = 10^{-5}M_{\text{Pl}}, \quad m = 10^{-6}M_{\text{Pl}}, \tag{9.45}$$

and substituting them into the expressions (9.43), (9.44), we arrive at constraints on the parameters g and λ:

$$\sqrt{g/\lambda} \gg 10^{-5}, \quad g/\sqrt{\lambda} \ll 10^{-6}. \tag{9.46}$$

This set of inequalities is not very burdensome. Thus, if $\lambda \sim 1$, then $10^{-10} \ll g \ll 10^{-6}$. The smallness of the parameters is related to the choice of small parameters in (9.45), which is conventional for inflationary models. Under such restrictions on the parameters, the phase fluctuations have a sharp peak at some e-folding number, which we have denoted by N_f.

The second method of suppressing large-scale fluctuations is more cardinal and consists of the following. Recall that the Mexican hat potential has been postulated from the beginning. Its shape has no other foundation than convenience and effectiveness. But let us also recall the notion of a "landscape" that appears in string theory as well as in the mechanisms of cascade reduction and random potentials, to be discussed further in this book. As a whole, the landscape implies that the shape of the scalar field potential depends on a number of random circumstances such as the initial values of the field. Then the shape of the potential may be arbitrary. Let us employ this and slightly change the shape of the potential (9.14). Using the landscape ideas, we consider one of the possible forms of the potential

$$V(\chi) = V_b(\chi) \cdot F(\chi),$$
$$F(\chi) = \frac{C^2}{(|\chi|^a + C)^2}, \quad C, a = \text{const.} \tag{9.47}$$

The function $F(\chi) \xrightarrow[\chi \to 0]{} 1$, so that $V(\chi) \xrightarrow[\chi \to 0]{} V_b(\chi)$, and the modified potential is close to the original form V_b from (9.17) while χ is small.

Since the modified potential may be made less steep by varying the parameters a and C, the time required for the field to reach its minimum may be accordingly increased.

The χ field dynamics is governed by the equation of motion

$$\frac{\partial^2 \chi(t)}{\partial t^2} + 3H \frac{\partial \chi(t)}{\partial t} + \frac{\partial V(\chi)}{\partial \chi} = 0 \tag{9.48}$$

where H is the Hubble parameter. This differential equation can be solved numerically.

The number of e-folds since the beginning of inflation is $N = H \cdot t$. It can be seen that in the modified potential the field reaches its minimum ($|\chi|_{\text{min}} = |f|/\sqrt{2}$) slower.

The magnitude of the baryon fluctuation is proportional to the fluctuation $\Delta\theta$ of the phase θ which depends on time as

$$\Delta\theta(t) = \frac{H}{2\pi \cdot \overline{|\chi(t)|}} \tag{9.49}$$

each e-folding, where $\overline{|\chi(t)|}$ is the mean absolute value of the field χ during that e-folding. Figure 9.2 shows the evolution of $\Delta\theta$ in the modified potential. Due to the inverse relation between $\Delta\theta$ and χ, the modified potential allows the value of $\Delta\theta$ to be significantly smaller during the first e-foldings as compared to its value after $\sim$ 10th e-foldings. This is exactly what is needed to suppress large scale baryon fluctuations.

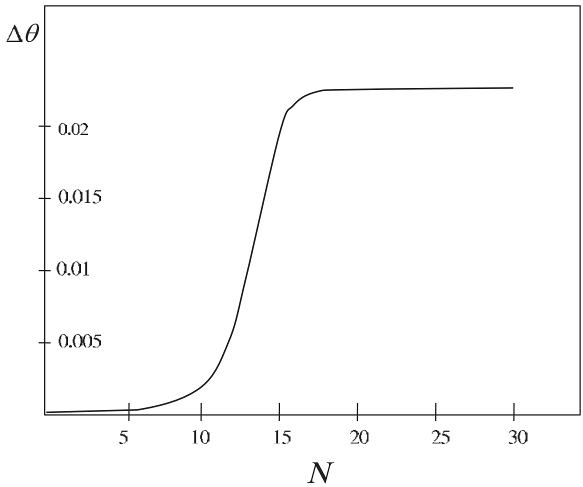


Figure 9.2 Computed values of $\Delta\theta$ as a function of N, the number of e-folds, $N = H \cdot t$. Parameter values: $H = 1, \lambda = 0.05, f = 10, a = 1.1, C = 10$.

So, in this chapter we have considered an example of an inflationary model with inhomogeneous baryogenesis: islands of antimatter, sufficiently spatially extended, emerge here in a natural way. Their number and size and the antimatter density inside them are controlled by a few parameters of the model.

Besides the model of baryogenesis, there exist a number of other possibilities, see, e.g., [6, 138]. It would be of interest to apply to them the ideas outlined here.

9.4 Massive primordial black holes

In the previous sections we have seen that quantum fluctuations during inflation can lead to such nontrivial phenomena as cosmologically large antimatter regions. Besides, while discussing the hybrid inflation model, it has been observed that inclusion of quantum fluctuations is able to cardinally change the picture of the system's further evolution.

Now we would like to discuss the role of quantum fluctuations in primordial black hole (PBH) formation. It is conventionally believed that the PBH mass is bounded above by a quantity of the order of a single solar mass. The suggested mechanism is of interest in that it allows for formation of massive and supermassive black holes long before the first stars appeared. Let us mention here in advance that this becomes possible because there appear closed domain walls whose size is much larger than that of the horizon right after inflation. While contracting, they deliver to the horizon an energy proportional to their surface area. We will consider a formation mechanism of such walls.

9.4.1 Field fluctuations near an extremum of the potential

Some inflationary models suppose creation of our Universe either near a maximum of the inflaton field potential or near its saddle point(s) to supply slow rolling during a sufficient number of e-folds. As will be shown below, these models include the possibility of formation of macroscopically large closed walls from a scalar field. After the end of inflation, these closed walls collapse to black holes if these walls are large and heavy enough [234, 356]. This mechanism is realized in well-known models like Hybrid Inflation [274] and Natural Inflation [139]. A scalar field could be the inflaton itself or some additional field.

First of all, we consider a general mechanism of closed wall formation based on quantum fluctuations near unstable point(s) like a saddle point or a maximum of the scalar field potential. An evolving scalar field may be split into a classical part, governed by the classical equation of motion, and quantum fluctuations [385].

Let us choose a positive value for the initial field, $\phi_{\text{in}} > 0$, as illustrated in Fig. 9.3. Then the mean field value will increase with time, ultimately reaching a minimum of the potential at some value $\phi_+ > 0$. It means that a greater part of space will be finally filled with the field value $\phi = \phi_+$. Meanwhile, the field in some (small) spatial domain could jump over the maximum due to quantum fluctuations. Later on, the mean value of the field representing this fluctuation tends to another minimum of the potential, $\phi_- < 0$. As a result, the space at the final stage will be filled with the vacuum ϕ_+ while some spatial domains will be characterized by the field value $\phi = \phi_- < 0$. If one starts to move from inside the domain to the outside, the path will begin from a spatial point with ϕ_- and end at a point with ϕ_+. Hence, the spatial path must contain a point with the

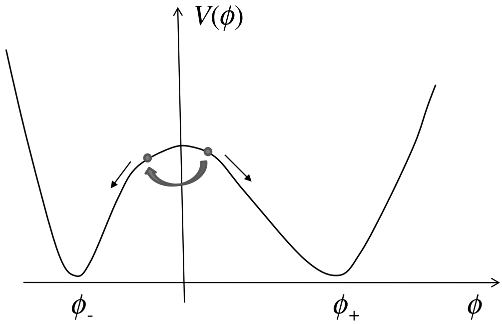


Figure 9.3 Quantum creation of walls during inflation. A right black point relates to the initial field value, $\phi_{\rm in}$.

maximum value of the potential. It means that a wall is inevitably formed between such spatial domain and the "outer" space with $\phi = \phi_+$.

So the "dangerous" values of fluctuations are those with $\phi \leq 0$. Such spatial domains will be surrounded by closed walls, and if their number is too large, it will strongly affect the dynamics of the early Universe. It is useful to calculate the probability of nucleation of these fluctuations. The latter depends on the initial field value $\phi_{\rm in}$ at the moment of creation of our Universe. The corresponding probability

$$P_0(\phi_{\rm in}, T) = \int_{\phi=-\infty}^{\phi=0} dP(\phi, T; \phi_{\rm in}, 0), \tag{9.50}$$

to find the field value $\phi \leq 0$ at some spatial point, for reasonable values of the parameters, is presented in Fig. 9.4. This probability determines the ratio of spatial volumes with different signs of the field.

To facilitate the analysis, let us approximate the potential near its maximum by

$$V = V_0 - \frac{m^2}{2}\phi^2, \tag{9.51}$$

where the maximum is assumed at $\phi = 0$ without loss of generality. Then the probability density to find a certain field value ϕ has the form [234] (adapted to the case under consideration):

$$dP(\phi, T; \phi_{\rm in}, 0) = d\phi \sqrt{\frac{a}{\pi(e^{2\mu T} - 1)}} \exp\left[-a\frac{(\phi - \phi_{\rm in} e^{\mu T})^2}{e^{2\mu T} - 1}\right]. \tag{9.52}$$

Here $a = \mu/\sigma^2$, $\mu \equiv m^2/(3H)$, and $\sigma = H^{3/2}/(2\pi)$, with the Hubble parameter $H \simeq \sqrt{8\pi V_0/(3M_{\rm Pl})}$.

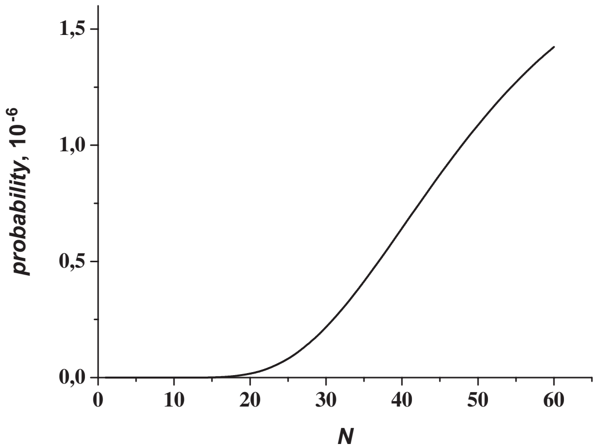


Figure 9.4 The part of space occupied by another vacuum state, depending on the time measured in e-folds. The initial field value is $\phi_{\text{in}} = \phi_{\max} + 3H$, where $\phi^{\max}$ is the field value at a maximum of the potential. The parameter m is chosen to be $m = 0.3H$, where H is the Hubble parameter at the top of the potential.

This probability is highly sensitive to the initial field value ϕ_{in}: the closer the nucleation to the maximum of the potential, the greater the part of the Universe that is covered with walls at the final stage.

If the fraction of space surrounded by the walls is not very large, the resulting massive black holes, formed from the walls, could explain the early formation of quasars.

The mass and space distribution of the resulting black holes strongly depends on the specific model and the parameters of the corresponding Lagrangian as well as on the initial conditions.

9.4.2 A specific example

Let us choose a specific form of the potential and study the possible PBH mass spectra. We start with a complex scalar field φ with the Lagrangian

$$L_{\varphi} = \frac{1}{2}|\partial\varphi|^2 - V(|\varphi|), \tag{9.53}$$

where $\varphi = r\,\mathrm{e}^{i\vartheta}$. The potential is chosen in the form (9.17)

$$V(|\varphi|) = \lambda(|\varphi|^2 - f^2/2)^2 + \Lambda^4(1 - \cos\vartheta), \quad \Lambda \ll f. \tag{9.54}$$

Here λ and f are parameters of the Lagrangian. This potential has already been used in this chapter, see Sec. 9.3. It has a countable set of saddle points at $\theta = \pi$ and is quite suitable for our goals.

We assume the mass of the radial field component r, i.e., the value

$$m_r^2 \equiv d^2V/dr^2|_{r=f/\sqrt{2}},$$

to be sufficiently large, so that the complex magnitude of the field acquires a value somewhere in the circular valley $|\varphi| \cong f/\sqrt{2}$ before the end of inflation. Since the minimum of the potential (9.54) is almost degenerate, the field has the form

$$\varphi \simeq f/\sqrt{2} \cdot \mathrm{e}^{i\vartheta(x)}. \tag{9.55}$$

In the framework of this model, the PBH spectra have been obtained and studied in detail in the book [234] and the article [357] where we refer the reader for details. The resulting spectra have a peculiar feature: an exponential growth of BH number as their mass decreases. A physical reason is evident: closer to the end of the inflationary period, more closed walls are formed and they are less "stretched" for the time remaining by the end of inflation. Recall that the mass of a wall is proportional to its area.

We see that this scenario, in addition to supermassive BH (SMBH) formation, also predicts the existence of intermediate-mass BHs (IMBHs) in galactic halos at significant distances from their centres. A distinguishing feature of this scenario is black hole formation even at the radiation-dominated stage as well as their spatial distribution in stellar clusters. While "astrophysical" black holes are formed in potential wells at centres of dark matter halos [133], the PBHs are, on the contrary, able to form around themselves an induced dark matter halo which looks like a dwarf galaxy [134, 135] with a sharp density growth towards the centre.

It would be wrong to imagine that absolutely all IMBHs merge to form central SMBHs in modern galaxies. It is more likely that among them only some fraction will merge while others should form an IMBH population in galactic halos. A similar situation emerges in many models with primordial SMBHs. Indeed, if a certain mechanism predicts the formation of a single SMBH in the volume of a galaxy, it is hard to avoid the presence of multiple less massive IMBHs in the same volume. Besides, even if the origin of SMBHs in galactic nuclei is not connected with IMBHs, the IMBH mass distribution function usually grows in the direction of small masses.

Thus a sufficiently common feature of the above-mentioned SMBH formation mechanisms is the predicted high abundance of IMBHs. However, the observational data substantially restrict the possible number of IMBHs in galaxies. Only a small number of objects might be attributed to IMBHs with a certain probability, e.g., IMBHs in some globular clusters, but there is so far no unambiguous proof of IMBH presence at their centres.

One has to state that the present observational data do not give a clear answer to the question of IMBH existence and abundance in the Universe. There are only particular examples and indications as well as theoretical arguments for their existence. In what follows we will modify the previous model in order to suppress IMBH formation in galaxies but preserve the effective SMBH formation in galactic nuclei. This variant can become topical if future observations substantially restrict the IMBH number.

9.4.3 Suppressed intermediate-mass black hole formation

Up to now we neglected the classical field evolution during inflation. Let us look what happens if we take it into account. For simplicity we will first assume that the field is moving along the valley of the potential $V(\vartheta)$. The latter means that Eq. (9.55) holds. Then the equation of motion for the field ϑ with the potential (9.54) reads

$$\ddot{\theta} + 3H\dot{\theta} + \frac{\Lambda^4}{f^2}\sin\theta = 0. \tag{9.56}$$

The Hubble parameter H remains constant during inflation since the field θ is not an inflaton. During inflation the quantity θ changes slowly, and only after its termination does θ rapidly oscillate near the minima, which is necessary for effective particle production and for heating of the Universe. Depending on the initial conditions, θ can roll down to one of the minima $\theta_{\min} = 0$ or $\theta_{\min} = 2\pi$.

The number and mass of PBHs which have formed as a final result strongly depend on the slope of the potential Λ and on the symmetry breaking scale f at the beginning of the inflation. In [136] such model parameters were chosen at which there form clusters initially consisting of massive PBHs, of which the largest ones reach the mass of $\sim 4 \cdot 10^7 M_\odot$. Such massive PBHs can eventually serve as protogalaxy cores, further increasing their mass up to $\sim 10^9 M_\odot$ due to accretion.

We are interested in the phase evolution with time due to both quantum fluctuations and classical motion. The suitable mathematical formalism was developed in [342, 385], where it was suggested to split the field into a sum of the classical component Θ and the fluctuations ϑ around it:

$$\theta = \Theta(t) + \vartheta, \quad \vartheta \ll \Theta. \tag{9.57}$$

The probability density $P_f(\vartheta)$ of finding the fluctuation part of the phase ϑ satisfies the Fokker–Planck equation whose solution has the form of a Gaussian distribution [385]. The phase distribution which is of interest for us is evidently found by the elementary substitution $P(\theta) = P_f(\Theta - \theta)$.

With the parameters chosen, $\Lambda = 1.75H$, $f = 10H$, the tilt of the potential (9.54) is small, and we will take it into account only when calculating the main, classical contribution $\Theta(t)$ in (9.57). Small corrections due to the fluctuations ϑ in (9.57) will be taken into account neglecting the slope of the potential in order not to exceed the required accuracy of the calculations. Meanwhile, in addition, for the spatial distribution of the phase, see analytical results, e.g., in [342, 385].

In terms of e-folds, the probability to find the phase in the interval $(\theta, \theta + \delta\theta)$ at a given spatial point is [234, 357]

$$\delta P(\theta) = \frac{1}{\sqrt{2\pi\,(N_r - N)}} \exp\left[-\frac{(\theta_r(N) - \theta)^2}{2\delta\theta^2(N_r - N)}\right]; \quad \delta\theta = \frac{H}{2\pi f}. \tag{9.58}$$

Here, as already mentioned, we have neglected the slope of the potential. The classical part of the phase $\Theta(t)$ is presented in the form $\theta_r(N)$, where the number of e-folds is $N = Ht$.

The condition that one and only one SMBH will appear at the instant N_0,

$$\delta P(\pi) \cdot e^{3(N_r - N_0)} = 1, \tag{9.59}$$

indicates that the BH is formed at crossing of the extremum point of the potential $\theta = \pi$. Now, from the equation

$$\mathrm{e}^{3(N_r - N_0)} = \sqrt{2\pi(N_r - N_0)} \exp\left[\frac{(\theta_r(N_0) - \pi)^2}{2\delta\theta^2(N_r - N_0)}\right], \tag{9.60}$$

we find the phase $\theta_r(N_0)$ with which the spatial region should be created such that at N_0 e-folds before the end of inflation the conditions for SMBH formation emerge:

$$\theta_r(N_0) = \pi - \delta\theta\sqrt{(N_r - N_0)[6(N_r - N_0) - \ln 2\pi(N_r - N_0)]}. \tag{9.61}$$

Let us now determine the probability of smaller-mass BH formation which takes place at the e-folding N, such that $N < N_0 < N_r$. Fig. 9.5 depicts the time dependence of the phase θ (where time is expressed in e-folds), obtained by numerically solving Eq. (9.56) and in the slow-rolling

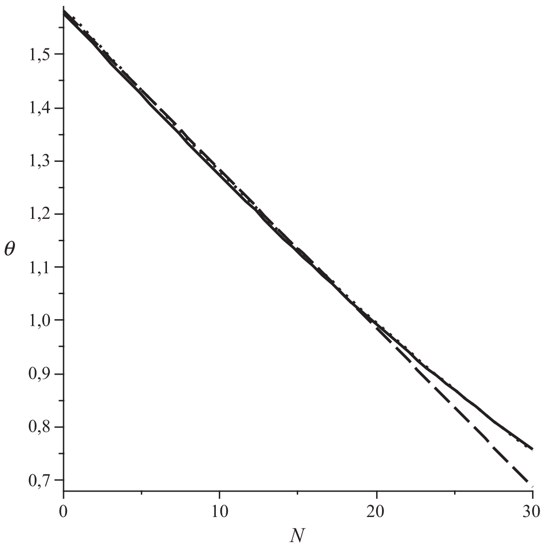


Figure 9.5 The dependence of θ on the number of e-folds at the values $\Lambda = 1.75H$, $f = 10H$. The solid line corresponds to a numerical calculation, dots (which almost merge with the line) denote the result of analytic calculation in the slow-rolling approximation. The approximating straight line is dashed, its slope coefficient is $\alpha = 0.028$.

approximation. The fine agreement shows that the system is indeed in a slow-rolling mode. In the subsequent calculation of the probability of quantum fluctuations of the field θ, it is expedient to use not the numerical solution of Eq. (9.56) but rather the linear dependence that approximates it. Taking into account that the primordial IMBHs are formed from large domains emerging at a small number of e-folds, the approximation of the solution to (9.56) is carried out in the initial part ($N < 20$). It is convenient to approximate the needed dependence by the linear function

$$\theta_r(N) \simeq \theta_r(N_0) - \alpha \cdot (N - N_0). \tag{9.62}$$

Then the probability of the conditions for IMBH formation emerging during the e-folding number N is

$$\delta P(\pi) = \frac{1}{\sqrt{2\pi(N_r - N)}} \exp\left[-\frac{(\theta_r(N_0) - \alpha \cdot (N - N_0) - \pi)^2}{2\delta\theta^2(N_r - N)}\right]. \tag{9.63}$$

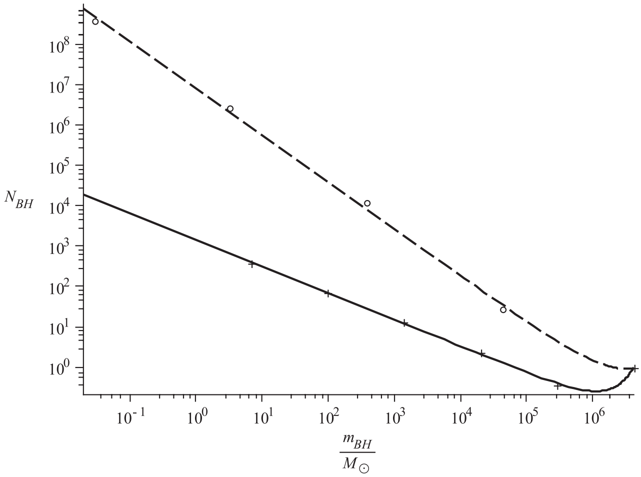


Figure 9.6 The PBH mass distribution. The solid line corresponds to the distribution for $\Lambda = 1.75H$, $f = 10H$ (the crosses mark the results of calculations), the dashed line to the distribution for $\alpha = 0$ (the calculated results are marked with circles).

Knowing the number of causally disconnected regions, $\exp[3(N - N_r)]$, which have formed since the formation time of the spatial region N_r, one can find the total IMBH number in this region. A detailed description has been given in [234], where, unlike our case, the classical field motion is not taken into account, i.e., $\alpha = 0$.

Figure 9.6 presents the PBH mass distribution for the parameters $\Lambda = 1.75H$, $f = 10H$. It is clear that the IMBH formation turns out to be really suppressed.

Thus the resulting BH mass spectrum is sensitive to the details of the classical motion and, in particular, to the initial field value. The logic and the relations obtained above will be useful in the next section.

9.4.4 PBH mass spectra and the scalar field dynamics

Let us look which results can be obtained in the general case assuming that the modern horizon has formed at $\varphi > f/\sqrt{2}$. For simplicity, we split the classical motion of the field $\Phi = \varphi(t)e^{i\theta(t)}$ in the potential (9.54) into two parts: the radial motion describing changes in the field magnitude,

$$\varphi(t), \quad \theta = \text{const},$$

when approaching the minimum $\varphi = f/\sqrt{2}$, and the transverse motion describing the evolution of the field phase in the valley of the potential:

$$\varphi = f/\sqrt{2}, \quad \theta = \theta(t).$$

During inflation, the space splits into a great number of mutually causally disconnected regions (but causally connected inside!), in which, due to quantum fluctuations, the field values are different. As described in [234], along the boundaries of regions with the phase $\theta = \pi$ there form closed field walls and, after their collapse, PBHs. The fluctuations of the field magnitude take place at radial motion of the field, which leads to changes in the BH mass spectrum.

As has been done above, the field is considered as a sum of the classical term Θ and the fluctuation ϑ around it,

$$\theta = \Theta(t) + \vartheta, \quad \vartheta \ll \Theta. \tag{9.64}$$

The probability to find the phase in the interval $(\pi, \pi + \delta\theta)$ by the end of inflation, under the condition that at some time the phase was $\theta = \theta(z)$, is

$$dP = \frac{1}{\sqrt{2\pi(z_u - z)}} \exp\left[-\frac{(\pi - \theta)^2}{2\delta\theta^2(z_u - z)}\right]. \tag{9.65}$$

Here $\delta\theta$ is the mean value of the phase fluctuation magnitude for a single e-folding, $\delta\theta = \frac{H}{2\pi f}$, z_u is the duration of the inflationary stage in which the observable Universe has formed, and z is the number of e-folds remaining to the end of inflation. In Eq. (9.65) it has been taken into account that the mean value $\Theta(t)$ of the phase θ is also time-dependent due to the classical motion. To include the field magnitude fluctuations during the separate motion, a similar formula is used:

$$dP(\varphi_r) = \frac{1}{\sqrt{2\pi(z_u - z)}} \exp\left[-\frac{(\varphi_r - \varphi)^2}{2\delta\varphi^2(z_u - z)}\right]. \tag{9.66}$$

Here $\delta\varphi = H/(2\pi)$ is the mean value of the fluctuation magnitude of the radial field component. The radial motion of the average field is described by the equation

$$\frac{\partial^2\varphi}{\partial t^2} + 3H\frac{\partial\varphi}{\partial t} + V'_\varphi = 0. \tag{9.67}$$

The form of the solution to (9.67) is mainly determined by the initial field value $\varphi_{\rm in} \neq f/\sqrt{2}$ $(\dot\varphi_{\rm in} = 0)$.

When the field magnitude reaches the value $\varphi \sim f/\sqrt{2}$, it is necessary to take into consideration the change in the field phase θ which, in this case, is described by the equation

$$\frac{\partial^2\theta}{\partial t^2} + 3H\frac{\partial\theta}{\partial t} + V'_\theta = 0. \tag{9.68}$$

The next step is to find the spatial size distribution of the domains with the phase $\theta > \pi$, taking into account the classical motion (9.67), (9.68). Such domains emerge due to quantum fluctuations of the scalar field.

The number of domains δN, filled with the phase $\theta > \pi$, emerging during a single e-folding at z e-folds before the end of inflation and creating a black hole in the future, is determined by the probability (9.65):

$$\delta N = N\,dP, \tag{9.69}$$

where the total number of domains N is $N = \mathrm{e}^{3z}$. By the end of inflation, these domains turn out to be surrounded by closed field walls. As a result of a collapse, almost the whole energy of such a closed wall can be concentrated in a small volume inside the gravitational radius, which is a necessary condition for BH formation.

The forming BH masses depend on the characteristic size of the closed wall, which is larger if a spatial domain surrounded by a given wall appeared earlier. The analytic dependence of the PBH mass on the wall size is discussed in [233, 234, 357]. Following these references and using the formulas of the present section for calculating the size distribution of the walls, we obtain the PBH mass spectrum.

Figure 9.7 shows the PBH spectrum in the observable Universe obtained from the above considerations. The initial field values $\varphi_{\text{in}} = 28H$, $\theta_{\text{in}} = 2.5$ lead to a sufficiently narrow spectrum. The maximum of this distribution is located at the BH mass of $\sim 10^{5-6} M_\odot$. Later on they can become SMBHs in galactic nuclei, having increased their mass due to matter accretion and merging with other BHs.

Less massive IMBHs form BH populations in galactic halos. It is also seen from Fig. 9.7 that, for instance, the number of IMBHs with the mass $\sim 10^3 M_\odot$ is suppressed by two orders of magnitude. That is, if the probability of SMBH formation in a certain galaxy is of order unity, then the probability of finding an IMBH is of the order of 1%. It can explain the difficulty of their search. The obtained PBH mass spectra are very sensitive to the initial values φ_{in} and θ_{in}.

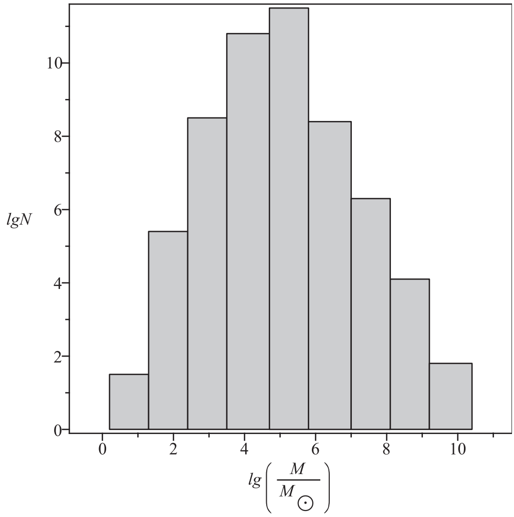


Figure 9.7 Primordial black hole mass distribution with the model parameters $\lambda = 0.005$, $\Lambda = 2.5H$, $f = 10H$, $\varphi_{\rm in} = 28H$, $\theta_{\rm in} = 2.5$.

9.4.5 Discussion

The galaxy formation process includes quantum creation of energy density fluctuations and their spatial growth at the inflationary stage, growth of the density after the fluctuation comes under the horizon, and the subsequent star formation. In the standard picture, the stars gradually accumulate near the centre and merge, thus forming a massive black hole whose mass grows with time. However, the possibility of galaxy formation with supermassive BHs at the centre is becoming more and more intriguing and unclear, especially in connection with the discovery of distant quasars with redshifts $z > 6$. Such an early formation of black holes with masses $\sim 10^9 M_\odot$ can become a serious problem for the standard models of BH formation [131]. Thus the scenarios with primordial massive BH formation [28, 100, 356, 435] attract attention as a possible alternative to the standard scenario. Such PBHs can be condensation centres for baryons and dark matter [131, 132] with subsequent protogalaxy formation.

In this chapter, we have described a PBH formation mechanism as a result of collapse of closed field walls as well as a formation mechanism of these walls themselves. It has been shown that the shape of the BH mass spectra strongly depends both on the parameters of the potential and on the initial conditions.

A further evolution of a closed wall goes on as follows. The wall possessing an energy excess tends to the most energetically favourable state. Like a soap bubble, it acquires a spherical shape and contracts. (If there were no air in a soap bubble, it would do the same). The fate of a small wall is of minor interest: it contracts to its centre, different parts of the wall collide with each other, and as a result all its energy passes on to expanding waves of the field which had formed the wall. The situation with large walls is quite different. Their contraction delivers enormous energy to a small spatial volume. This energy is held by the gravitational field, and a compact object is formed. A more detailed analysis shows that such objects, black holes, are formed in whole families, or clusters. It is evident that no stars are required to form these massive primordial black holes! On the contrary, it is such a BH cluster that can create a gravitational potential well and collect there a sufficient number of baryons from which stars can later form. There already exist observational data on solitary supermassive black holes, whose environment is free of luminous matter. This can be considered as an indirect confirmation of the above-described mechanism of massive PBH formation.

A particular realization of this mechanism has been developed in the papers [132, 234, 235, 354, 356]. In [132] such model parameters have been chosen that from the very beginning there form large clusters of massive PBHs with central BH masses of $\sim 4 \cdot 10^7 M_\odot$, which grow by accretion up to values $\sim 10^9 M_\odot$ and thus describe the observed activity of the distant quasars.

In this scenario, there are several stages of BH and galaxy formation:

1. Formation of closed walls of a scalar field immediately after the end of inflation and their collapse to a PBH cluster according to [235, 356, 357]. The most massive PBH is formed at the centre of the cluster after crossing the cosmological horizon.

2. Separation of the central dense region of the cluster from the cosmological expansion and its virialization. A large number of less massive PBHs that surround the central one merge with it and increase its mass.

3. Separation of the external regions of the cluster, dominated by dark matter, from the cosmological expansion, and growth of a protogalaxy. End of protogalaxy growth due to interaction with ordinary fluctuations of dark matter.

4. Gas cooling and star formation, merging of protogalaxies and formation of modern galaxies.

One cannot exclude the possibility that both galaxy formation scenarios, the standard one and the one with PBHs, are realized simultaneously. With all that, since at the initial stage the galaxies often merge, it is sufficient that a small fraction of protogalaxies possess massive black holes at their centres to explain the observed abundance of supermassive black holes at galactic centres.

Part III

Extra Dimensions

This page intentionally left blank

The idea of extra dimensions, tracing back to the pioneering works of Thomas Kaluza [222] and Oscar Klein [241], became in the last decades a necessary component of almost all attempts to unify the four physical interactions. The most popular are models with dimensions $D = 10$ (superstrings) and 11 (M-theory), but there are quite a number of still higher-dimensional models. The bosonic sector of such theories usually contains, apart from the metric, different kinds of scalar and vector fields as well as antisymmetric form fields; meanwhile, it is well known that, in the extra (with respect to the usual four) dimensions, the D-dimensional metric components naturally create scalar and vector fields when reduced to the four-dimensional space-time (see, e.g., [407] and references therein).

It is the simplest case of such a reduction, leading to Einstein's gravity and Maxwell's electromagnetic field, that has been described in Kaluza's paper [222]. More complex unified multidimensional theories, providing the advent of boson fields that appear in the Standard Model of particle physics from multidimensional gravity, have been described in the book [407] (see there also references to other works in this area, including numerous papers by members of Yu.S. Vladimirov's group). A separate problem to be solved is geometrization of fermion fields.

Of special interest are multidimensional theories with a gravitational action containing terms nonlinear with respect to curvature. Their emergence directly follows from quantum field theory in curved space-time [38, 178], and therefore it may not even be considered as an independent postulate.

According to [77, 83], multidimensional gravity with curvature-nonlinear Lagrangians leads to a diversity of low-energy theories owing to arbitrariness of the input parameters of the gravitational action and different variants of the initial data. The latter can include both the metric and the topology of extra dimensions that can emerge from quantum fluctuations at high energies. Thus even at fixed parameters of the initial effective Lagrangian, one can expect the emergence of many kinds of Universes with drastically different properties. Even under comparatively simple assumptions on the nature of multidimensional geometry, after reduction to the four observed dimensions there appear effective scalar fields with complicated potentials, whose properties to a large extent determine the physical properties of one or another universe.

Such a viewpoint is close to the concept of chaotic inflation, according to which there permanently emerge indefinitely many universes at fluctuations of a scalar field (inflaton). With a sufficiently sophisticated shape of the potential, the situation resembles the so-called landscape concept in string theory: the number of different vacua in heterotic string theory reaches 10^{1500} [269], the number of more realistic de Sitter vacua [220] is also huge though finite. We try to make a next step, explaining the origin of such potentials from the standpoint of multidimensional gravity.

Obtaining particular results in multidimensional gravity is substantially simplified by using the slow-change approximation [77, 83] that assumes small values of all derivatives (in a sense to be discussed below) and smallness of all energy densities as compared to Planck scales. Such assumptions turn out to hold even in early cosmology characterized by Grand Unification energies, to say nothing about processes at lower energies and in the realm of modern cosmology. This method allows for considering quite different kinds of gravitational actions and the wide ranges of parameters they contain.

In the Kaluza–Klein concept, the extra spatial dimensions remain invisible due to their extreme smallness. In the recent years, however, an alternative concept became a subject of active studies, namely, the brane-world concept, which considers our Universe as a distinguished three-dimensional (or four-dimensional if time is included) surface or layer, called the brane, in a multidimensional space-time where the extra dimensions have large or even infinite size. They remain invisible because the Standard Model fields (and hence the observers) are assumed to be concentrated on the brane while gravity (and as a rule only gravity) is allowed to propagate in the surrounding space, called the bulk. The history of such models is traced back to the early 80s, with papers by Akama [9], Rubakov and Shaposhnikov [351] and others. The outburst of interest in such models is basically related to some achievements of string theory and M-theory, in particular, with the well-known Hořava–Witten 11-dimensional model [208, 429], in which one of the extra dimensions has a much larger size than others. This approach was shown to be able to suggest a natural mechanism of solving the hierarchy problem in particle physics ([336] and others), while in weak fields the Newtonian law of gravity is preserved as required by modern experiments. Other unsolved basic problems of particle physics as well as cosmology have received significant progress in brane-world models.

A separate problem is the very number of extra dimensions which varies from zero to infinity in works by different authors. Introducing the notion of "extended superspace", we will arbitrarily vary this number. Originally, the concept of superspace meant the set of various geometries [423]; subsequently, the set of all possible topologies was included in it. Let us take one more step and extend the superspace to include spaces of various dimensionalities. To be more precise, let us define the extended superspace E as a direct product of superspaces $\mathbb{M}$ of various dimensionalities:

$$\mathrm{E} = \mathbb{M}_1 \times \mathbb{M}_2 \times \mathbb{M}_3 \times \cdots \times \mathbb{M}_D \ldots$$

Here, $\mathbb{M}_n$ is a superspace of dimension $n = 1, 2, \ldots, D$, which is the set of all possible geometries (up to diffeomorphisms) and topologies.

Quantum fluctuations generate various geometries in each of the superspaces (space-time foam) [423]. The probability of a quantum birth of "long-lived" 3-geometries and the conditions under which this occurs are discussed below. In this chapter, we consider the corollaries of the hypothesis on the existence of an extended superspace.

This page intentionally left blank

Chapter 10

Multidimensional gravity

In this chapter, we will describe a few classes of model universes originating in nonlinear multidimensional gravity, without invoking any fields other than the metric one in the initial action. Among them are cosmological models able to describe both inflation in the early Universe and the modern stage of accelerated expansion.

10.1 Compact extra dimensions. A brief review

The idea that our world can have a dimensionality greater than four has appeared at the beginning of last century. The idea has turned out to be both simple and deep. Simple because it is rather simple to imagine, at least using two- and three-dimensional analogues. Thus, a sheet of paper is perceived (and used) as a two-dimensional object, but for microbes it is unambiguously three-dimensional. Evidently, due to a small size of, say, the fifth dimension it can avoid our perception. The idea is deep owing to its numerous and unexpected consequences. To obtain significant results, it is necessary to master the corresponding mathematical tools. Therefore in what follows we briefly present the necessary information.

Consider a D-dimensional differentiable manifold $\mathbb{M}_D$, which can be presented as a direct product,

$$\mathbb{M}_D = \mathbb{M}_4 \times \mathbb{M}_d. \tag{10.1}$$

The extra-dimensional space $\mathbb{M}_d$ is supposed to be compact and have a volume $V_d < \infty$. Let us introduce some coordinates in $\mathbb{M}_D$:

$\{X\} = (\{x\}, \{y\})$, where the set x^μ, $\mu = 1, 2, 3, 4$ describes the four-dimensional space-time $\mathbb{M}_4$ and the set y_a, $a = 1, 2, \ldots, d$ the extra space $\mathbb{M}_d$. The interval is written in terms of the metric tensor as usual:

$$ds^2 = G_{MN} dX^M dX^N, \tag{10.2}$$

where the indices M, N run over the values $1, 2, \ldots, d + 4$. The volume of the extra space is expressed in terms of the positive-definite "internal" metric γ_{ab}:

$$V_d = V(\mathbb{M}_d) = \int d^d y \sqrt{\gamma} < \infty, \tag{10.3}$$

where $\gamma = \det(\gamma_{ab})$. The latter inequality is a necessary condition of compactness of this subspace.

Scalar fields and extra spaces

To demonstrate the opportunities, let us consider the behaviour of a scalar field Φ defined in the full space $\mathbb{M}_D$, with the action

$$S = \int_{\mathbb{M}_D} d^D X \sqrt{|G|} \left[\frac{1}{2} (\partial_M \Phi \partial^M \Phi) - U(\Phi) \right]. \tag{10.4}$$

Varying the action in Φ with a fixed metric, we obtain the equation of motion

$$\Box_D \Phi + dU/d\Phi = 0.$$

Here, the D-dimensional d'Alembert operator is defined as follows:

$$\Box_D \Phi \equiv \nabla_M \nabla^M \Phi = \frac{1}{\sqrt{|G|}} \partial_M (\sqrt{|G|} \partial^M \Phi), \tag{10.5}$$

where $G = \det(G_{MN})$. If we restrict ourselves to small deflections from an equilibrium position (a minimum of the potential), we can approximate the potential with a quadratic expression $U(\Phi) \simeq M^2 \Phi^2 / 2$. Then the equation of motion simplifies to take the form

$$(\Box_D + M^2)\Phi = (\Box_4 + \Box_d + M^2)\Phi = 0,$$

where $\Box_4$ and $\Box_d$ are d'Alembertian operators in the corresponding subspaces.

Let us introduce an orthonormal set of functions $Y_n(y)$, such that

$$\Box_d Y_n(y) = \lambda_n Y_n(y). \tag{10.6}$$

Since $\mathbb{M}_d$ is a compact space, the eigenvalues of the d'Alembert operator λ_n are discrete. The orthonormality condition is written as follows:

$$\int_{\mathbb{M}_d} d^d y \sqrt{|\gamma|} Y_n(y)^* Y_m(y) = \delta_{nm} \tag{10.7}$$

Expanding Φ in a series in the eigenfunctions $Y_n(y)$,

$$\Phi(X) = \sum_n Y_n(y)\Phi_n(x), \tag{10.8}$$

and taking into account the expressions (10.6) and (10.7), it is easy to write the Φ field equation in the space $\mathbb{M}_4$:

$$(\Box_4 + \lambda_n + M^2)\Phi_n(x) = 0. \tag{10.9}$$

The fields $\Phi_n(x)$ are perceived by an observer as scalar fields with masses $M_n = M^2 + \lambda_n$. The nonobservability of scalar particles appearing from extra dimensions is usually explained by their large masses.

Let us consider a widespread example of extra space. Suppose that the extra space $\mathbb{M}_d$ has the geometry of a d-dimensional torus, $\mathbb{M}_d = \mathbb{T}^d \times \mathbb{S}^1 \times S^1 \times S^1 \times \ldots \times S^1$, $y^m \in [0,\ 2\pi R]$. Then the explicit form of the eigenvalues (10.6) is well known:

$$\lambda_n \equiv \lambda_{\{n_1,n_2,\ldots,n_d\}} = \sum_{j=1}^{d} \frac{n_j^2}{R^2}. \tag{10.10}$$

The eigenfunctions then have the form

$$Y_{\{n_1,n_2,\ldots,n_d\}}(y) = \frac{1}{\sqrt{V_T}} \exp\left[\frac{i\sum n_j y^j}{R}\right], \quad V_T = (2\pi R)^d. \tag{10.11}$$

If the radius R of the torus, characterizing the size of the extra space, is small, then all excitations with $j \geq 1$ correspond to supermassive particles.

10.1.1 A Kaluza–Klein model with a single extra dimension

A closed one-dimensional manifold (a closed curve) is, from the viewpoint of its internal geometry, a circle $\mathbb{S}^1$. Therefore the structure (10.1) in the case $d = 1$ is

$$\mathbb{M}_5 = \mathbb{M}_4 \times \mathbb{S}^1. \tag{10.12}$$

It is convenient to represent the metric of the full space $\mathbb{M}_5$ as a matrix:

$$G_{AB} = \begin{pmatrix} G_{\mu\nu} & G_{\mu 5} \\ G_{5\mu} & G_{55} \end{pmatrix}, \tag{10.13}$$

where G_{55}, $G_{5\mu}$, $G_{\mu\nu}$ are the scalar, vector and tensor in the 4-dimensional space $\mathbb{M}_4$.

How to extract gauge symmetry from extra dimensions

To understand the idea, let us make the restriction $G_{55} = \text{const} = -1$. Then transformations of the form

$$\begin{aligned} x'^{\mu} &= x^{\mu}, \\ y' &= y + \varepsilon(x), \quad \varepsilon(x) \ll 1 \end{aligned} \tag{10.14}$$

affect the metric in the following way:

$$G'_{\mu\nu} = \frac{\partial x^A}{\partial x'^{\mu}} \frac{\partial x^B}{\partial x'^{\nu}} G_{AB}, \quad G'_{\mu 5} = \frac{\partial x^A}{\partial x'^{\mu}} G_{A5},$$

while the coefficient G_{55} remains invariable. In our case of infinitesimal shifts,

$$\begin{aligned} G'_{\mu\nu} &= G_{\mu\nu} - G_{\nu 5}\partial_{\mu}\varepsilon - G_{\mu 5}\partial_{\nu}\varepsilon + O(\varepsilon^2), \\ G'_{\mu 5} &= G_{\mu 5} - G_{55}\partial_{\mu}\varepsilon = G_{\mu 5} + \partial_{\mu}\varepsilon. \end{aligned} \tag{10.15}$$

The very idea of extra dimensions has emerged from a wish to describe the gravitational field and the electromagnetic (gauge) field in a unified form. As components of the latter, one can take the off-diagonal elements of the metric tensor $G_{\mu 5}$, because, as is seen from (10.15), their transformations strongly resemble gauge ones. The situation is somewhat more complex with the choice of the four-dimensional metric. A choice of $G_{\mu\nu}$ as the metric tensor in the main subspace $\mathbb{M}_4$ does not seem good because it depends on the choice of the extra-space coordinate, (10.15). On the other hand, it is not hard to understand that

$$g_{\mu\nu} = G_{\mu\nu} + G_{\mu 5}G_{\nu 5},$$

is invariant under the infinitesimal transformations (10.14). It is this tensor that is convenient as a choice of the metric tensor in the main subspace $\mathbb{M}_4$. Denoting $B_{\mu} = -G_{\mu 5}$, we write the metric tensor in the form

$$G_{AB}(x) = \begin{pmatrix} g_{\mu\nu} - B_{\mu}B_{\nu} & -B_{\mu} \\ -B_{\nu} & -1 \end{pmatrix}. \tag{10.16}$$

The inverse matrix of G_{AB} is

$$G^{AB}(x) = \begin{pmatrix} g^{\mu\nu} & -B^{\mu} \\ -B^{\nu} & -1 + B_{\lambda}B^{\lambda} \end{pmatrix}. \tag{10.17}$$

The quantity B_μ, under the transformation (10.14), is transformed as

$$B'_\mu = B_\mu - \partial_\mu \varepsilon, \tag{10.18}$$

which is of particular interest: it is in this way that the vector potential of the electromagnetic field transforms under the action of the group $U(1)$ on the wave function of a charged particle. The greatest advantage of Kaluza–Klein theories is that gauge symmetries follow from the symmetry of a compact extra space instead of being additionally postulated.

The invariance of the action (10.4) under the transformations (10.14) causes the emergence of gauge invariance of this action under the transformations (10.18).

Let us look how the transformation (10.14) affects a real scalar field $\Phi(x, y)$. It can be represented by the Fourier series (10.8)

$$\Phi(x, y) = \sum_n \Phi_n(x) Y_n(y),$$

where $Y_n(y)$ is an orthonormalized set of eigenfunctions of the operator ∂_y^2 acting in the extra space $\mathbb{M}_1$. Since in our case it is a circle of a certain radius r, we have

$$Y_n(y) = \frac{1}{\sqrt{2\pi r}} \, \mathrm{e}^{-iny/r}.$$

The symmetry of the extra space (circle) manifests itself in the invariance of the functions $Y_n(y)$ under the shifts $y \to y + 2\pi r$. The scalar field is transformed by the substitution $y \to y + \varepsilon$ in the following way:

$$\Phi'(x, y) = \Phi(x, y + \varepsilon).$$

Comparing the Fourier transforms of both parts of the equality, we obtain

$$\Phi'_n(x) = \Phi_n(x) e^{i(n/r)\varepsilon}.$$

It is in this way that a charged scalar field transforms at gauge transformations. It means that the Fourier component $\Phi_n(x)$ can be perceived as a scalar field with the charge $q_n \equiv n/r$.

We see that $U(1)$ symmetry of the extra space implies the existence of the gauge 4-vector B_μ and charged particles in the space $\mathbb{M}_4$. It is clear that the Lagrangian in four dimensions that we are going to obtain will be invariant under gauge transformations since they are created by the transformations (10.14), under which the initial Lagrangian is manifestly invariant. Nevertheless, to obtain an explicit form of the low-energy action, some further reasoning is necessary.

Let the action of an arbitrary scalar field Φ have the form

$$S = \int d^5 z \sqrt{G} \left(-\frac{1}{2} \Phi \Box_G \Phi - \frac{1}{2} m^2 \Phi^2 \right).$$

The purely kinetic term in the action is written as

$$\frac{1}{2} \frac{\partial \Phi}{\partial x^A} \frac{\partial \Phi}{\partial x_A}.$$

The two expressions differ by a total derivative that does not affect the classical equations of motion for the field Φ. We choose the five-dimensional metric in the simplest form and approximate it as follows, see (10.16):

$$G_{AB}(x) = \begin{pmatrix} \eta_{\mu\nu} - B_\mu B_\nu & -B_\mu \\ -B_\nu & -1 \end{pmatrix}. \tag{10.19}$$

We here neglect the dependence of the gauge field $B_\mu(x)$ on the coordinate y of the extra space and the dynamics of the metric coefficient G_{55}. It means physically that, by assumption, all relaxation processes in the fifth dimension have finished. The d'Alembert operator is presented as

$$\Box_G = \eta^{\mu\nu}(\partial_\mu - B_\mu \partial_y)(\partial_\nu - B_\nu \partial_y) + \Box_y.$$

Substituting this expression into the action, we obtain

$$S = \frac{1}{2} \sum_n \int d^4 x \sqrt{-g} \Phi_n^* (-\Box_n + m_n^2) \Phi_n,$$

where

$$\Box_n = g^{\mu\nu}(\partial_\mu - i q_n A_\mu)(\partial_\nu - i q_n A_\nu), \quad A_\mu \equiv \frac{-1}{\sqrt{2\varkappa}} B_\mu,$$
$$q_n = n(\sqrt{2\varkappa}/R), \quad m_n^2 = m^2 + n^2/R^2.$$

In the language of field theory, the resulting effective action describes a set of scalar fields with masses m_n^2 and charges q_n, interacting with the gauge field A_μ in Minkowski space.

Generalization: $G_{55} \neq const$

Up to now we have been considering the simplest version of a 5D theory where the metric coefficient G_{55} is coordinate-independent. Let us extend the boundaries of what was presented by assuming $G_{55} \neq \text{const}$.

It is convenient to introduce the field B_μ, generalizing the one introduced previously in the following way:

$$G_{\mu 5} = -\sqrt{|G_{55}|}B_\mu. \tag{10.20}$$

As will be shown a little later, the 5-metric can be written as

$$\begin{pmatrix} g_{\mu\nu} - e^{2\phi}B_\mu B_\nu & -\mathrm{e}^{\phi}B_\mu \\ -\mathrm{e}^{\phi}B_\mu & -e^{2\phi} \end{pmatrix}, \tag{10.21}$$

where we have introduced the scalar field ϕ:

$$G_{55} \equiv -e^{2\phi}. \tag{10.22}$$

This result makes it possible to single out the gauge symmetry in the 5-dimensional Hilbert–Einstein action. Indeed, the Ricci scalar R_5 of the 5D space is expressed in terms of the corresponding scalar R_4 of the 4D space as follows:

$$R_5 = R_4 - 2e^{-\phi}\Box e^{\phi} - \frac{1}{4}e^{2\phi}F_{\mu\nu}F^{\mu\nu}; \tag{10.23}$$

$$F_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu. \tag{10.24}$$

Then the 5D Einstein-Hilbert action, after some calculation, takes the form

$$\begin{aligned} S &= \frac{1}{16\pi G_5}\int_{\mathbb{M}_5} d^5X\sqrt{|G|}(R_5 - 2\Lambda) \\ &= \frac{1}{16\pi G_4}\int_{\mathbb{M}_4} d^4x\sqrt{|g|}\left(R_4 - 2\Lambda' - \frac{3}{2}\partial_\mu\phi\partial^\mu\phi - \frac{1}{4}\,\mathrm{e}^{3\phi}F_{\mu\nu}F^{\mu\nu}\right). \end{aligned} \tag{10.25}$$

The derivation of this formula is omitted due to its lengthiness. During the calculation, the following conformal transformation was done:

$$g_{\mu\nu} \to \tilde{g}_{\mu\nu} = \mathrm{e}^{\phi}g_{\mu\nu}$$

and we have denoted $G_4 = G_5/(2\pi R)$, $\Lambda' = \frac{4}{3}\Lambda$. In (10.25) the metric $\tilde{g}$ is used.

Thus, having assumed the existence of a single extra dimension, we have obtained 4D gravity with a massless scalar field interacting with an Abelian gauge (electromagnetic) field. The theory with such an interaction is often called dilaton gravity, and the field ϕ is called the dilaton. If $\phi = \mathrm{const}$, it is the usual Einstein–Maxwell theory.

The geometric meaning of $g_{\mu\nu}$

The extra space in the form of a circle has the symmetry (isometry group) $U(1)$, whose transformations are rotations by an arbitrary angle. The metric tensor $g_{\mu\nu}$ of the main (i.e., our) space is invariant under such transformations. It is this group that acts as a gauge group for the gauge field B_μ obtained from the geometry and interpreted as the electromagnetic field.

Owing to the simplicity of the extra space we were able to easily construct a suitable combination from the quantities $G_{\mu\nu}$, $G_{\mu 5}$ and G_{55}. In more involved cases this procedure is not so obvious. It is therefore necessary to come to a clearer understanding of the meaning of the quantities introduced.

Since the metric tensor G_{AB} has off-diagonal elements, the space $\mathbb{M}_4$ with the metric $G_{\mu\nu}$ is not orthogonal to $\mathbb{M}_1$. Let us find a space $\mathbb{U}_4$ orthogonal to $\mathbb{M}_1$. We will seek a small displacement vector, orthogonal to $\mathbb{M}_1$, in the form

$$\vec{\delta} \equiv dx^\mu \vec{e}_\mu + dy\vec{e}_5, \tag{10.26}$$

where $\vec{e}_\mu$, $\vec{e}_5$ are a set of five unit vectors, and $\vec{e}_5$ is tangent to $\mathbb{M}_1$. For the orthogonality condition written as

$$(dx^\mu \vec{e}_\mu + dgy\vec{e}_5) \cdot \vec{e}_5 = 0,$$

we find the displacement dy:

$$dy = -G_{55}^{-1} G_{5\mu} dx^\mu.$$

The absolute value of the displacement vector is

$$\vec{\delta}^2 = \delta^M G_{MN} \delta^N = (G_{\mu\nu} - G_{55}^{-1} G_{\mu 5} G_{\nu 5}) dx^\mu dx^\nu \equiv g_{\mu\nu} dx^\mu dx^\nu. \tag{10.27}$$

Here we have used the equalities $\delta^\mu = dx^\mu, \quad \delta^5 = dy$, see (10.26). Now the meaning of the tensor $g_{\mu\nu}$ is clear: it determines the interval in a space orthogonal to $\mathbb{M}_1$.

Vielbeins

Let us introduce the vielbein formalism which is convenient for extension of the above results to compact extra spaces of d spatial dimensions.

It is well known that gravity can be described by a full set (basis) of linearly independent vielbein vectors e_A instead of the metric tensor. The index A is the number of the vector in $\mathbb{M}_5$, $A = 1, 2, \ldots, 5$. The metric tensor components g_{AB} are defined in terms of the vectors e_A using their scalar product:

$$g_{AB} = e_A \cdot e_B.$$

The basis vectors can be chosen in such a way that the metric tensor defined with their aid have the canonical form $\eta_{AB} = \mathrm{diag}(1, -1, -1, \ldots, -1)$. The vielbein consisting of the vectors $e_{\bar{A}}$ such that $e_{\bar{A}} e_{\bar{B}} = \eta_{AB}$ is called the Lorentz basis. One basis can be expanded by another:

$$e_A = b_A^{\bar{A}} e_{\bar{A}} = b_A^{\bar{\mu}} e_{\bar{\mu}} + b_A^{\bar{5}} e_{\bar{5}}. \tag{10.28}$$

We assume summing over repeated indices. The vectors e_5 and $e_{\bar{5}}$ are chosen to be collinear, $e_5 = b_5^{\bar{5}} e_{\bar{5}}$. Then Eq. (10.28) immediately implies that $b_5^{\bar{\mu}} = 0$. The expansion (10.28) can be presented in a matrix form, with the matrix

$$\begin{pmatrix} b_\mu^{\bar{\mu}} & 0 \\ b_\mu^{\bar{5}} & b_5^{\bar{5}} \end{pmatrix},$$

acting on the columns $(e_{\bar{\mu}}, e_{\bar{5}}), e_\mu, e_5$. The inverse matrix

$$\begin{pmatrix} h_{\bar{\mu}}^{\mu} & h_{\bar{\mu}}^{5} \\ h_{\bar{5}}^{\mu} & h_{\bar{5}}^{5} \end{pmatrix}$$

has an important property that

$$h_{\bar{5}}^{5} b_5^{\bar{5}} = 1.$$

The components of the metric tensor in $\mathbb{M}_4$ are expressed as follows:

$$\begin{aligned} G_{\mu\nu} &\equiv e_\mu \cdot e_\nu = b_\mu^{\bar{A}} b_\nu^{\bar{B}} e_{\bar{A}} e_{\bar{B}} = b_\mu^{\bar{\mu}} b_\nu^{\bar{\nu}} e_{\bar{\mu}} e_{\bar{\nu}} + b_\mu^{\bar{5}} b_\nu^{\bar{5}} e_{\bar{5}} e_{\bar{5}} \\ &= b_\mu^{\bar{\mu}} b_\nu^{\bar{\nu}} \eta_{\bar{\mu}\bar{\nu}} + b_\mu^{\bar{5}} b_\nu^{\bar{5}} \eta_{\bar{5}\bar{5}} = g_{\mu\nu} - b_\mu^{\bar{5}} b_\nu^{\bar{5}}, \end{aligned} \tag{10.29}$$

where we have introduced the tensor $g_{\mu\nu} \equiv b_\mu^{\bar{\mu}} b_\nu^{\bar{\nu}} e_{\bar{\mu}} e_{\bar{\nu}} = b_\mu^{\bar{\mu}} b_\nu^{\bar{\nu}} \eta_{\bar{\mu}\bar{\nu}}$, which is obviously invariant under coordinate transformations in the extra space. Besides,

$$\begin{aligned} G_{\mu 5} &\equiv e_\mu \cdot e_5 = b_\mu^{\bar{A}} b_5^{\bar{B}} e_{\bar{A}} e_{\bar{B}} = b_\mu^{\bar{5}} b_5^{\bar{5}} e_{\bar{5}} e_{\bar{5}} = -b_\mu^{\bar{5}} b_5^{\bar{5}}, \\ G_{55} &\equiv e_5 \cdot e_5 = b_5^{\bar{5}} b_5^{\bar{5}} e_{\bar{5}} e_{\bar{5}} = -b_5^{\bar{5}} b_5^{\bar{5}}. \end{aligned}$$

Defining the field B_μ by

$$B_\mu \equiv B_\mu^5 \equiv h_{\bar{5}}^5 b_\mu^{\bar{5}},$$

we arrive at the expressions obtained above (see (10.21), (10.22), where $b_5^{\bar{5}} = e^\phi$):

$$\begin{aligned} G_{\mu\nu} &= g_{\mu\nu} - b_5^{\bar{5}} b_5^{\bar{5}} B_\mu B_\nu; \\ G_{\mu 5} &= -b_5^{\bar{5}} B_\mu; \\ G_{55} &= -b_5^{\bar{5}} b_5^{\bar{5}}. \end{aligned}$$

10.1.2 Kaluza–Klein models. The general case

There is no reason for which it would be necessary to restrict oneself to a single extra dimension. As will be seen, an increased number of dimensions yields rich opportunities. Let us therefore consider a Riemannian space with an arbitrary number of dimensions, singling out, for evident reasons, the four-dimensional physical subspace.

In the previous section, we have considered splitting of the metric tensor using the vielbein method. This metric is readily extended to an arbitrary number of extra dimensions. In what follows we do things quite similar to the 5D variant.

We introduce the basis vectors: the coordinate basis e_A for splitting $\mathbb{M}_4 \times \mathbb{M}_d$, and $e_{\bar{A}}$, the Lorentz basis of the same space-time. One basis can be expanded by the other:

$$e_A = b_A^{\bar{A}} e_{\bar{A}} = b_A^{\bar{\mu}} e_{\bar{\mu}} + b_A^{\bar{a}} e_{\bar{a}}, \tag{10.30}$$

where again summing is assumed. The indices μ, ν refer to the space $\mathbb{M}_4$ while $a, b, \ldots$ to the space $\mathbb{M}_d$.

The expansion (10.30), by full analogy with the previous case, can be represented in a matrix form with the transition matrix

$$\begin{pmatrix} b_\mu^{\bar{\mu}} & b_a^{\bar{\mu}} \\ b_\mu^{\bar{a}} & b_a^{\bar{a}} \end{pmatrix}.$$

In the Lorentz basis, the scalar product $e_{\bar{\mu}} \cdot e_{\bar{a}} = 0$ by definition. Consequently, $b_a^{\bar{\mu}} = 0$, while the matrix

$$\begin{pmatrix} h_{\bar{\mu}}^{\mu} & h_{\bar{\mu}}^{a} \\ h_{\bar{a}}^{\mu} & h_{\bar{a}}^{a} \end{pmatrix},$$

is the most general form of the inverse matrix.

The components of the metric tensor G in $\mathbb{M}_4$ can be expressed in terms of the expansion coefficients of one basis with respect to the other:

$$\begin{aligned} G_{\mu\nu} &\equiv e_\mu \cdot e_\nu = b_\mu^{\bar{A}} b_\nu^{\bar{C}} e_{\bar{A}} e_{\bar{C}} = b_\mu^{\bar{\mu}} b_\nu^{\bar{\nu}} e_{\bar{\mu}} e_{\bar{\nu}} + b_\mu^{\bar{a}} b_\nu^{\bar{c}} e_{\bar{a}} e_{\bar{c}} \\ &= b_\mu^{\bar{\mu}} b_\nu^{\bar{\nu}} \eta_{\bar{\mu}\bar{\nu}} + b_\mu^{\bar{a}} b_\nu^{\bar{c}} \eta_{\bar{a}\bar{c}} = g_{\mu\nu} - b_\mu^{\bar{a}} b_\nu^{\bar{a}}. \end{aligned} \tag{10.31}$$

Besides,

$$G_{\mu a} \equiv e_\mu \cdot e_a = b_\mu^{\bar{A}} b_a^{\bar{C}} e_{\bar{A}} e_{\bar{C}} = b_\mu^{\bar{a}} b_a^{\bar{c}} e_{\bar{a}} e_{\bar{c}} = -b_\mu^{\bar{a}} b_a^{\bar{a}}, \tag{10.32}$$

and

$$G_{ac} \equiv e_a \cdot e_c = b_a^{\bar{a}} b_c^{\bar{c}} e_{\bar{a}} e_{\bar{c}} = -b_a^{\bar{a}} b_c^{\bar{a}}. \tag{10.33}$$

Here we have used the basic property of the Lorentzian basis that $e_{\bar{A}} e_{\bar{C}} = \eta_{\bar{A}\bar{C}}$. Let us define the field B_μ as follows:

$$b^{\bar{a}}_\mu = b^{\bar{a}}_a B^a_\mu,$$

or explicitly

$$B_\mu \equiv B^a_\mu \equiv h^a_{\bar{a}} b^{\bar{a}}_\mu.$$

It addition, it is useful to denote

$$G_{ac} \equiv \phi_{ac},$$

thus reminding the reader that this expression is a scalar with respect to transformations in M_4, Substituting this definition into (10.31), (10.32), (10.33), we obtain the desired formulae for splitting the metric tensor of multidimensional space:

$$G_{\mu\nu} = g_{\mu\nu} + \phi_{ac} B^a_\mu B^c_\nu, \tag{10.34}$$

$$G_{\mu a} = \phi_{ac} B^c_\mu. \tag{10.35}$$

Let us give an explicit form of coordinate transformations for these quantities. Thus, a transformation similar to the 5D case

$$x' = x'(x), \quad y' = y'(x, y),$$

leads to the following relations:

$$g'_{\mu\nu} = \frac{\partial x^\lambda}{\partial x'^\mu} \frac{\partial x^\rho}{\partial x'^\nu} g_{\mu\nu}, \qquad G'_{ac} = \frac{\partial y^d}{\partial y'^a} \frac{\partial y^f}{\partial y'^c} G_{df}, \tag{10.36}$$

$$B'_\mu = \frac{\partial x^\nu}{\partial x'^\mu} \frac{\partial y'^a}{\partial y^c} B^c_\nu + \frac{\partial y'^a}{\partial y^c} \frac{\partial y^c}{\partial x^\nu}. \tag{10.37}$$

All these relations are valid without taking into account the symmetry of the extra space. Previously, when studying the 5D model, one could ignore the issue of form invariance of the extra space with respect to the transformation (10.14). Due to the simple geometry of the space $\mathbb{M}_1$, the form of its extra-space metric did not change under the transformations (10.14). If, however, the extra space is multidimensional, then arbitrary coordinate transformations, in general, affect the metric. For this reason, in the low-energy limit, the Lagrangian will not have the standard gauge-invariant form, although the off-diagonal components of the metric will be associated with vector gauge fields.

To obtain a gauge-invariant theory at low energies, it is necessary that the extra space possess the corresponding symmetry. Then the more narrow class of transformations applied to the metric of this space, the one that leaves the metric form-invariant, creates a gauge-invariant theory. We will see what are the consequences of this situation.

Killing vectors

As was previously shown, local transformations of the extra space $\mathbb{M}_d$ only,

$$y^a \to y'^a = y^a + \epsilon^a(x),$$

can be considered as transformations of a gauge field, at least in the 5D case. In general, this affects both the metric G_{AB} of the main space $\mathbb{M}_4$ and that of the extra space $\mathbb{M}_d$. While in the main space it is possible to single out the invariant part $g_{\mu\nu}$, transformations of the extra space metric is an undesirable feature. Therefore it makes sense to address coordinate transformations that leave the extra space metric invariable.

A coordinate transformation $y \to y'$ is called an isometry if it leaves the metric form-invariant:

$$G_{ab}(y) \to G'_{ab}(y') = G_{ab}(y'). \tag{10.38}$$

If such coordinate transformations form a group, it is called an isometry group $\mathcal{G}(\mathbb{M}_d)$ of the space $\mathbb{M}_d$. We will be interested in infinitesimal transformations.

Let there be a set of vectors $k_i^m(y)$, $i = 1, 2, \ldots, r$ such that small shifts along them

$$y^m \to y'^m = y^m + k_i^m(y)\varepsilon^i(x), \quad \varepsilon \ll 1 \tag{10.39}$$

leave the metric of $\mathbb{M}_d$ invariable. The vectors $k_i^m(y)$ are called Killing vectors and are generators of the isometry of this space. It can be shown that these vectors satisfy the commutation relations

$$k_i^a \partial_a k_j^b - k_j^a \partial_a k_i^b = f_{ij}^l k_l^b. \tag{10.40}$$

The structure constants f_{ij}^l are determined by the specific isometry group. The maximum possible number of independent Killing vectors is related to the dimension of space in the following way:

$$\max A = \frac{d(d+1)}{2}. \tag{10.41}$$

The equations for determining the Killing vectors can be obtained from (10.38) and read

$$\nabla_a k_{ib} + \nabla_b k_{ia} = 0, \quad k_{ia} \equiv G_{ab} k_i^b(y). \tag{10.42}$$

They are called the Killing equations.

A non-Abelian gauge group as an isometry group of the extra space

Let us apply the above equations for studying a relationship between the symmetry of the extra space and gauge invariance of a low-energy theory in four dimensions. We choose the following coordinate dependence of the metric:

$$g_{\mu\nu} = g_{\mu\nu}(x), \quad G_{mn} = G_{mn}(y). \tag{10.43}$$

Furthermore, we expand the vector B_μ^m in a basis of Killing vectors k_i^n:

$$B_\mu^m = k_i^m(y) B_\mu^i(x). \tag{10.44}$$

This can be done because B_μ^m is a vector in the space $\mathbb{M}_d$, possessing the corresponding symmetry.

Now we can show that the isometry group in $\mathbb{M}_d$ is realized in $\mathbb{M}_4$ as a non-Abelian gauge group. To this end we consider the infinitesimal transformations

$$x'^\mu = x^\mu, \qquad y'^\alpha = y^\alpha + k_i^\alpha(y)\varepsilon^i(x).$$

After some calculations, from the transformations (10.36) follows the transformation law for the object B_μ^i:

$$\delta B_\mu^i = -\partial_\mu \varepsilon^i + f_{jl}^i B_\mu^j \varepsilon^l.$$

It is precisely what is called a gauge transformation of the group G. The extra space metric $G_{ab}(y)$ is invariant under these transformations because k_α^i are Killing vectors of the extra space.

The standard gravitational action in D-dimensional space is

$$S_G = \frac{1}{2\varkappa_D} \int d^D X \sqrt{G}(R_D - 2\Lambda). \tag{10.45}$$

Lengthy calculations lead to the following expression for the Ricci scalar:

$$R_D = R_4 + R_d - \frac{1}{4} G_{ab} k_i^a k_j^b F_{\mu\nu}^i F^{j\mu\nu}, \tag{10.46}$$

$$F_{\mu\nu}^i = \partial_\mu B_\nu^i - \partial_\nu B_\mu^i - f_{jl}^i B_\mu^j B_\nu^l, \quad F^{j\mu\nu} = g^{\mu\lambda} g^{\nu\rho} F_{\lambda\rho}^i. \tag{10.47}$$

Substituting this expression into the action, we find

$$S_G = \frac{1}{2\varkappa}\int d^4x d^D y\sqrt{-g}\sqrt{G_d}\left(R_4 + R_d - \frac{1}{4}G_{ab}k_i^a k_j^b F_{\mu\nu}^i F^{j\mu\nu} - 2\Lambda\right).$$

Let us now use our assumption that the extra space metric $G_{ab}(y)$ does not depend on x. For that reason, one can normalize the Killing vectors as follows:

$$\frac{1}{2\varkappa}\int d^d y\sqrt{G}G_{ab}k_i^a(y)k_j^b(y) = \delta_{ij}.$$

Thus the purely gravitational action in D-dimensional space describes gravity and a gauge field in 4D space:

$$S = \int d^4x\sqrt{-g}\left[-\frac{1}{2\varkappa}R_4 - \frac{1}{4}F_{\mu\nu}^i(x)F^{i\mu\nu}(x) - 2\Lambda'\right]. \qquad (10.48)$$

10.2 Multidimensional gravity with higher-order derivatives. Basic equations

In what follows we develop an approach, suggested by the authors, which allows for reducing multidimensional models with higher-order derivatives to the 4D Einstein–Hilbert action with scalar fields.

10.2.1 $F(R)$-theory

The Einstein–Hilbert action for a gravitational field, linear in curvature R, completely describes the physical phenomena at low energies where gravity is important. However, it is clear that quantum effects inevitably lead to nonlinear corrections in the expression for the action [143]. In this case, the action should contain terms with higher derivatives in the form of polynomials of various degrees in the Ricci scalar and other invariants. Thus, whatever gravitational action we take as the basis, after applying quantum corrections it takes the form

$$S = \int d^N x(R + \varepsilon_1 R^2 + \varepsilon_2 R^3 + \varepsilon_3 R^4 + \cdots + \alpha_1 R_{AB}R^{AB} + \cdots),$$

with a set of unknown coefficients depending on the topology of space [95, 345, 387]. However, the problem is not that acute, since the nonlinear (in Ricci scalar) theories can be reduced to a linear theory by a conformal transformation. Moreover, in our papers [77, 83], we have suggested a more general method for reducing arbitrary Lagrangians to the standard

Einstein–Hilbert form in the low-energy limit. In this case, the problem of stabilizing the size of the extra dimensions [101] turns out to be quite solvable.

Consider a $(D = d_0 + d_1)$-dimensional manifold $\mathbb{M}$, having the simplest geometric structure of a direct product, $\mathbb{M} = \mathbb{M}_0 \times \mathbb{M}_1$, with the metric

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu + \mathrm{e}^{2\beta(x)} b_{ab} dx^a dx^b, \tag{10.49}$$

where the extra-space metric components b_{ab} do not depend on x^μ, the observable space-time coordinates. For realistic models $d_0 = 4$, but in Eqs. (10.49)–(10.53) we leave an arbitrary d_0 for generality.

The D-dimensional Riemann tensor has the following nonzero components:

$$\begin{aligned} R^{\mu\nu}{}_{\rho\sigma} &= \overline{R}^{\mu\nu}{}_{\rho\sigma}, \\ R^{\mu a}{}_{\nu b} &= \delta^a_b\, B^\mu_\nu, \quad B^\mu_\nu := \mathrm{e}^{-\beta} \nabla_\nu (\mathrm{e}^\beta \beta^\mu), \\ R^{ab}{}_{cd} &= \mathrm{e}^{-2\beta} \overline{R}^{ab}{}_{cd} + \delta^{ab}{}_{cd} \beta_\mu \beta^\mu, \end{aligned} \tag{10.50}$$

where capital Latin indices run over all D values, an overbar marks quantities obtained separately from $g_{\mu\nu}$ and b_{ab}, $\beta_\mu \equiv \partial_\mu \beta$, and $\delta^{ab}{}_{cd} \equiv \delta^a_c \delta^b_d - \delta^a_d \delta^b_c$. The nonzero components of the Ricci tensor and the scalar curvature read

$$\begin{aligned} R^\nu_\mu &= \overline{R}^\nu_\mu + d_1\, B^\nu_\mu, \\ R^b_a &= \mathrm{e}^{-2\beta} \overline{R}^b_a + \delta^b_a [\Box\beta + d_1 (\partial\beta)^2], \\ R &= \overline{R}[g] + \mathrm{e}^{-2\beta} \overline{R}[b] + 2 d_1 \Box\beta + d_1 (d_1 + 1)(\partial\beta)^2, \end{aligned} \tag{10.51}$$

where $(\partial\beta)^2 \equiv \beta_\mu \beta^\mu$, $\Box = g^{\mu\nu} \nabla_\mu \nabla_\nu$ is the d_0-dimensional d'Alembert operator, and $\overline{R}[g]$ and $\overline{R}[b]$ are the Ricci scalars obtained from $g_{\mu\nu}$ and b_{ab}, respectively.

Suppose that b_{ab} describes a compact d_1-dimensional space of nonzero constant curvature, i.e., a sphere $(k = 1)$ or a compact d_1-dimensional hyperbolic space[20] $(k = -1)$ with a fixed curvature radius r_0 normalized to the D-dimensional analogue m_D of the Planck mass, i.e., $r_0 = 1/m_D$

[20]Compact hyperbolic spaces of constant curvature on the basis of the conventional Lobachevsky space $\mathbb{H}^d$ are isometric to $\mathbb{H}^d/\Gamma$, where Γ is a nontrivial discrete subgroup of the isometry group of $\mathbb{H}^d$, see, e.g.., [146]. On possible applications of such (3D) spaces in cosmology see, e.g., [305] and references therein.

(here, as everywhere, we use the natural units, with the speed of light c and Planck's constant $\hbar$ equal to unity). We have

$$\begin{aligned} \overline{R}^{ab}{}_{cd} &= k\, m_D^2\, \delta^{ab}{}_{cd}, \\ \overline{R}_a^b &= k\, m_D^2\, (d_1 - 1)\delta_a^b, \\ \overline{R}[b] &= k\, m_D^2\, d_1(d_1 - 1) = R_b. \end{aligned} \tag{10.52}$$

The scale factor $b(x) \equiv \mathrm{e}^\beta$ in (10.49) is thus kept dimensionless; R_b has the meaning of a characteristic curvature scale of the extra dimensions.

Consider, in the above geometry, a sufficiently general curvature-nonlinear theory of gravity with the action

$$S = \frac{1}{2} m_D^{D-2} \int \sqrt{{}^D g}\, d^D x\, [F(R) + L_m], \tag{10.53}$$

where $F(R)$ is an arbitrary smooth function, L_m is a matter Lagrangian and ${}^D g = |\det(g_{MN})|$. The extra coordinates are easily integrated out, and the action is reduced to $d_0 = 4$ dimensions:

$$S = \frac{1}{2} \mathcal{V}[d_1]\, m_D^2 \int \sqrt{{}^4 g}\, d^4 x\; \mathrm{e}^{d_1 \beta}\, [F(R) + L_m], \tag{10.54}$$

where ${}^4 g = |\det(g_{\mu\nu})|$ and $\mathcal{V}[d_1]$ is the volume of a compact d_1-dimensional space of unit curvature.

10.2.2 Slow-change approximation. The Einstein frame

Equation (10.54) describes a 4D theory which is nonlinear in curvature and, moreover, contains a nonminimal coupling between the effective scalar field β and the curvature. A significant progress can be achieved by using an expansion in the smallness parameter $L_d/L_D \ll 1$, where L_d and L_D are the characteristic sizes of the extra and basic spaces. We will also consider only such geometries whose Ricci scalars satisfy the condition

$$R_d \gg R_D. \tag{10.55}$$

These two conditions do not contradict each other for such widely used geometries as, e.g., a d-dimensional sphere.

Let us simplify the theory in the following manner:

(a) Express everything in terms of 4-dimensional variables and $\beta(x)$; note that now

$$R = R_4 + \phi + f_1,$$
$$R_4 = \overline{R}[g], \quad f_1 = 2d_1\Box\beta + d_1(d_1+1)(\partial\beta)^2, \tag{10.56}$$

and we have introduced the effective scalar field

$$\phi(x) = R_b\, \mathrm{e}^{-2\beta(x)} = kd_1(d_1-1)m_D^2\, \mathrm{e}^{-2\beta(x)}. \tag{10.57}$$

Recall that we have $k = \pm 1$ for positive and negative curvature in d_1 extra dimensions, respectively, so that ϕ has different signs in these cases by definition.

(b) Suppose that all quantities are slowly varying, i.e., consider each derivative ∂_μ (including those in the definition of $\overline{R}$) as an expression containing a small parameter ε; neglect all quantities of orders higher than $O(\varepsilon^2)$ (see [143]).

(c) Perform a conformal mapping leading to the Einstein conformal frame, where the 4-curvature appears to be minimally coupled to the scalar ϕ.

In the decomposition (10.56), both terms f_1 and R_4 are regarded small in our approach, which actually means that all quantities, including the 4D curvature, are small as compared to the D-dimensional Planck scale. So the only term which is not small is ϕ, and we can use a Taylor decomposition of the function $F(R) = F(\phi + R_4 + f_1)$:

$$F(R) = F(\phi + R_4 + f_1) \simeq F(\phi) + F'(\phi)\cdot(R_4 + f_1) + \cdots, \tag{10.58}$$

with $F'(\phi) \equiv dF/d\phi$. Substituting it into Eq. (10.54), we obtain up to $O(\varepsilon^2)$

$$S = \frac{1}{2}\mathcal{V}[d_1]\, m_D^2 \int \sqrt{{}^4g}\, d^4x\; \mathrm{e}^{d_1\beta}[F'(\phi)R_4 + F(\phi) + F'(\phi)f_1 + L_m], \tag{10.59}$$

where β is related to ϕ according to (10.57). The expression (10.59) is typical of a scalar-tensor theory (STT) of gravity in a Jordan frame.

For an analysis of the scalar field dynamics and, in particular, for finding stationary points, it is helpful to pass on to the Einstein frame where ϕ is minimally coupled to gravity. After the conformal mapping

$$g_{\mu\nu} \mapsto \widetilde{g}_{\mu\nu} = |f(\phi)| g_{\mu\nu}, \quad f(\phi) = \mathrm{e}^{d_1\beta} F'(\phi), \tag{10.60}$$

with the corresponding transformation of the scalar curvature

$$R_4 = |f| \left(\widetilde{R}_4 + \frac{3}{f}\widetilde{\Box} f - \frac{3}{2f^2}(\widetilde{\partial} f)^2 \right), \tag{10.61}$$

(the tilde marks quantities obtained from or with $\widetilde{g}_{\mu\nu}$) the action (10.59) acquires the form

$$S = \frac{1}{2}\mathcal{V}[d_1]\, m_D^2 \int \sqrt{\widetilde{g}}\, d^4x \biggl\{ [\operatorname{sign} F'(\phi)]\ [\widetilde{R}_4 + K(\phi)] - V(\phi) + \frac{\mathrm{e}^{-d_1\beta}}{F'(\phi)^2} L_m \biggr\}, \tag{10.62}$$

with the kinetic (K) and potential (V) terms

$$\begin{aligned} K(\phi) &= \frac{3}{2}\frac{(\partial f)^2}{f^2} - 2d_1 \frac{f^{,\mu}\beta_\mu}{f} + (\partial\beta)^2, \\ V(\phi) &= -\,\mathrm{e}^{-d_1\beta}\, F(\phi)/F'(\phi)^2. \end{aligned} \tag{10.63}$$

It remains to express everything in terms of a single scalar variable, say, ϕ. We can write the action (10.62) in the form

$$\begin{aligned} S &= \frac{\mathcal{V}[d_1]}{2} m_D^2 \int d^4x \sqrt{\widetilde{g}}\, (\operatorname{sign} F') L, \\ L &= \widetilde{R}_4 + \frac{1}{2} K_{\rm Ein}(\phi)(\partial\phi)^2 - V_{\rm Ein}(\phi) + \widetilde{L}_m, \end{aligned} \tag{10.64}$$

$$\widetilde{L}_m = (\operatorname{sign} F') \frac{\mathrm{e}^{-d_1\beta}}{F'(\phi)^2} L_m; \tag{10.65}$$

$$K_{\rm Ein}(\phi) = \frac{1}{2\phi^2}\left[6\phi^2 \left(\frac{F''}{F'}\right)^2 - 2d_1\phi\frac{F''}{F'} + \frac{1}{2}d_1(d_1{+}2) \right], \tag{10.66}$$

$$V_{\rm Ein}(\phi) = -(\operatorname{sign} F') \left[\frac{|\phi| m_D^{-2}}{d_1(d_1-1)} \right]^{d_1/2} \frac{F(\phi)}{F'(\phi)^2}. \tag{10.67}$$

In (10.61)–(10.64), the indices are raised and lowered with the metric $\widetilde{g}_{\mu\nu}$; everywhere $F = F(\phi)$ and $F' = dF/d\phi$.

Equations (10.64)–(10.67) are valid for both positive ($\phi > 0$) and negative ($\phi < 0$) curvature of the extra dimensions. Minima of the potential $V_{\rm Ein}$, defined in the Einstein frame, determine stationary points of the scalar field, and they remain stationary points after a transition to any other conformal frame, including the original Jordan frame. It is the Einstein frame that determines the scalar field behavior since its dynamics near an extremum is governed directly by the potential $V(\phi)$ and only implicitly by the metric.

If ϕ resides at a minimum of the potential $V_{\rm Ein}$, this potential turns into an effective cosmological constant. Being applied to the present cosmological epoch, it can determine the observable dark energy density that drives the accelerated expansion of the Universe. Alternatively, in principle, it may drive inflation in the early Universe.

The potential (10.67) valid for any function $F(R)$ looks quite complex to have some nontrivial extrema. In addition, we have obtained rather a complex form of the kinetic term (10.66). Its properties are known to be as important for the field dynamics as the shape of the potential. For example, as shown in Ref. [260], zeros and singular points of the kinetic term may describe stable states of a scalar field.

In our expressions, both the effective potential (10.67) and the kinetic term (10.66) are singular at the values of ϕ where $f(\phi) \sim F'(R)$ [the factor before R_4 in the Jordan-frame action (10.59)] is zero. Unlike many papers restricted to $F'(R) > 0$, we also include models with $F'(R) < 0$. As will be seen below, this opens new promising possibilities such as new minima of the effective potential at which the extra dimensions may be stabilized.

In (10.64)–(10.67) we have changed the sign of the Lagrangian in the case $F' < 0$; to preserve the attractive nature of gravity for ordinary matter, the matter Lagrangian density should appear with an unusual sign from the beginning. As a result, the sign of the whole action of gravity and matter will be unusual. However, there is no effect on the matter equations of motion, and the conventional form of the effective Einstein equations at a stationary value of ϕ will also be preserved. Only the action of the ϕ field itself can be unusual, according to Eqs. (10.66) and (10.67). It should also be noted that the common sign of the total action does not affect quantum transitions as well. Indeed, the transition amplitude is expressed in the path integral technique as $\int \exp(iS[q])Dq$ where $q(t)$ is some dynamic variable. The transition probability

$$\iint \exp(iS[q_1] - iS[q_2]) Dq_1 Dq_2,$$

is invariant under the substitution $i \to -i$ (with interchanging the integration variables q_1 and q_2).

The following remark is in order here. There is a well-known conformal transformation [25, 26, 148, 156, 228, 285, 287, 428] that leads the action (10.53) to a form characteristic of the Einstein theory with a minimally coupled scalar field χ (instead of an additional degree of freedom connected with R), whose potential is expressed in terms of the function $F(R)$. This is a D-dimensional Einstein picture. The theory thus obtained can be further reduced by one more conformal transformation, now depending on $\beta(x)$, to the 4D Einstein picture. This has been done, in particular, in the papers [189, 190] for quadratic gravity. Our method is technically simpler and is more universal since (see below) it is able to work with other invariants as well, not only $F(R)$. In cases where our approximation works, the results of the two approaches coincide, as shown in [83].

Some estimates

To confront the theory with reality, it is necessary to choose a physical (observational) conformal picture. Let us suppose that the original Jordan frame in which the theory has been formulated coincides with the observational frame in which measurement standards do not change in space and time. A general discussion of the problem of choosing the physical conformal frame can be found in [82].

Thus we use the Einstein frame only as a convenient tool for finding stationary states.

In any observational frame, in the expected stationary state, the gravitational action has the approximate Einstein–Hilbert form

$$S_{\rm GR} = \frac{1}{2} m_4^2 \int \sqrt{{}^4 g}\, d^4 x\, (R_4^* - 2\Lambda_{\rm eff}), \qquad (10.68)$$

where $m_4 = (8\pi G_N)^{-1/2}$ is the 4D Planck mass, G_N is the effective Newtonian constant and R_4^* is the observable 4D curvature (for our choice, $R_4^* = R_4$). To be consistent with our approach and with observational constraints, this stationary state should satisfy the following requirements:

(i) A classical space-time description should be admissible, i.e., the true size of the extra dimensions should exceed the true D-dimensional Planck length $1/m_D$, the fundamental length scale of the theory:

$$m_D b_0 = e^{\beta_0} \gg 1, \quad \Leftrightarrow \quad \phi/m_D^2 \ll d_1(d_1 - 1). \qquad (10.69)$$

(ii) The slow-change approximation should work, which, for $\phi = \text{const}$, reduces to the requirement $R_4 \ll m_D^2$.

(iii) The observed size b^* of the extra dimensions can be much larger than the Planck length $l_4 = 1/m_4 \approx 8{\cdot}10^{-33}$ cm, but not larger than about 10^{-17} cm, which corresponds to the TeV energy scale.

(iv) The predicted effective cosmological constant $\Lambda_{\rm eff}$ should be very small to conform to the observations:

$$\Lambda_{\rm eff}/m_4^2 \sim 10^{-120}. \tag{10.70}$$

Obtaining such a small value as (10.70) is one of the well-known problems of theoretical cosmology since it is hard to explain without fine tuning why Λm_4^2, generally associated with vacuum energy density, is so many orders of magnitude smaller than the characteristic energy densities inherent to the known physical interactions (e.g., the Planck density m_4^4 for gravity).

Thus we assume that observations are performed in the Jordan frame. The corresponding action (10.59) is approximated by (10.68) with the effective constants

$$\frac{1}{8\pi G_N} = m_4^2 = \mathcal{V}[d_1]\, m_D^{D-2}\, b_0^{d_1} F_0', \quad \Lambda_{\rm eff} = -\frac{F(\phi_0)}{2F_0'}, \tag{10.71}$$

where $b_0 = {\rm e}^{\beta(\phi_0)}/m_D = b_*$ is the observable size of the extra dimensions and $F_0' = F'(\phi_0)$.

Eq. (10.71) leads to the following relation between the dimensionless quantities $b_0 m_D$ and $b_0 m_4$:

$$b_0^2 m_4^2 = \mathcal{V}[d_1]\, F_0'\, (b_0 m_D)^{d_1+2}. \tag{10.72}$$

The factor $\mathcal{V}[d_1]$ is of order unity; the same may be expected from the dimensionless quantity F_0'. By item (iii) above, $b_0^2 m_4^2 \lesssim 10^{30}$, and Eq. (10.72) gives

$$\begin{aligned} &b_0 m_D \lesssim 10^{30/(d_1+2)},\\ &m_4/m_D \sim (b_0 m_D)^{d_1} \lesssim 10^{15 d_1/(d_1+2)}. \end{aligned} \tag{10.73}$$

Thus m_D does not too much differ from m_4. It means, in particular, that our slow-change approximation (item (ii) above) works manifestly well in almost all thinkable circumstances since R_4 is the observable curvature. If m_4/m_D is sufficiently close to unity, this approximation is even valid for the curvature characteristic of primordial inflation at the Grand Unification scale, such that $R_4/m_4^2 \sim 10^{-6}$ [83]. In particular models this inequality is strengthened, to say nothing of the modern stage of the evolution.

Let us return to the inequalities (10.73) and the estimate of $\Lambda_{\rm eff}$. The ratio $\Lambda_{\rm eff}/m_4^2$ can be expressed as follows:

$$
\begin{aligned}
-\frac{\Lambda_{\rm eff}}{m_4^2} &= \frac{F(\phi_0)}{2\,m_D^2 F_0'}\frac{m_D^2 b_0^2}{m_4^2 b_0^2}\\
&= \frac{1}{2F_0'}\,\frac{F(\phi_0)}{m_D^2}(F_0'\mathcal{V}[d_1])^{\frac{-2}{d_1+2}}(m_4 b_0)^{\frac{-2d_1}{d_1+2}}. \qquad (10.74)
\end{aligned}
$$

If we try to make this ratio small, the last factor in (10.74) can give at best about 10^{-30} (for sufficiently large d_1). The remaining 90 orders of magnitude must be gained due to unnatural smallness of the dimensionless quantity $F(\phi_0)/m_D^2$ (actually, of the initial cosmological constant) and/or greatness of F_0'. The latter variant is, however, unsuitable since it will make $b_0 m_D \gg 1$ impossible, see (10.72).

As a result, the problem of fine tuning remains topical, though it should be noted that the very small value of 10^{-30} appears in this approach without any artificial effort.

10.2.3 The first generalization: A more general form of the Lagrangian

The method described above allows for considering wider classes of Lagrangians. Let us demonstrate that by adding terms proportional to the Ricci tensor squared $R_{AB}R^{AB}$ and the Kretschmann scalar $\mathcal{K} = R_{ABCD}R^{ABCD}$. By common views, these and other high-order curvature terms appear due to quantum corrections, and it seems natural to include them on equal footing with cR^2. Now the action has the form

$$
S = \frac{\mathcal{V}[d_1]}{2\varkappa^2}\int d^4x\,\sqrt{{}^4g}\,\mathrm{e}^{d_1\beta}(R + cR^2 + c_1 R_{AB}R^{AB} + c_2\mathcal{K} - 2\Lambda), \qquad (10.75)
$$

where the internal variables have been integrated out in full analogy with Eq. (10.53). For the metric (10.49), it is easy to obtain expressions for $R_{AB}R^{AB}$ and $\mathcal{K}$:

$$
\begin{aligned}
R_{AB}R^{AB} &= \overline{R}_{\mu\nu}\overline{R}^{\mu\nu} + 2d_1\overline{R}_{\mu\nu}B^{\mu\nu} + \mathrm{e}^{-4\beta}\overline{R}_{ab}\overline{R}^{ab}\\
&\quad + 2\,\mathrm{e}^{-2\beta}\overline{R}[h][\Box\beta + d_1(\partial\beta)^2] + d_1[\Box\beta + d_1(\partial\beta)^2]^2, \qquad (10.76)\\
\mathcal{K} &= \overline{\mathcal{K}}[g] + 4d_1 B_{\mu\nu}B^{\mu\nu} + \mathrm{e}^{-4\beta}\overline{\mathcal{K}}[h]\\
&\quad + 4\,\mathrm{e}^{-2\beta}\overline{R}[h](\partial\beta)^2 + 2d_1(d_1-1)[(\partial\beta)^2]^2. \qquad (10.77)
\end{aligned}
$$

In the slow-change approximation, in the same manner as before, we obtain the 4D effective Lagrangian

$$\sqrt{{}^4g}L = \sqrt{{}^4g}\, \mathrm{e}^{d_1\beta}[R_4(1+2c\phi) + \phi + c_{\mathrm{tot}}\phi^2 - 2\Lambda + (1+2c\phi)f_1 \\ + 2c_1\phi\Box\beta + 2(c_1d_1 + 2c_2)(\partial\beta)^2], \tag{10.78}$$

where

$$c_{\mathrm{tot}} = c + \frac{c_1}{d_1} + \frac{2c_2}{d_1(d_1-1)}.$$

The conformal mapping (10.60) leads to the Einstein-frame Lagrangian (10.64) with the kinetic and potential terms

$$K_{\mathrm{Ein}}(\phi) = K_{\mathrm{Ein}}^{(2)}(\phi) + \frac{c_1+c_2}{2\phi(1+2c\phi)}, \tag{10.79}$$

$$V_{\mathrm{Ein}}(\phi) = -\,\mathrm{sign}(1+2c\phi) \\ \times \left[\frac{|\phi|}{d_1(d_1-1)}\right]^{d_1/2} \frac{c_{\mathrm{tot}}\phi^2 + \phi - 2\Lambda}{(1+2c\phi)^2}, \tag{10.80}$$

where the term $K_{\mathrm{Ein}}^{(2)}(\phi)$ is taken from (10.66).

The presence of the parameters c_1 and c_2 adds freedom in choosing the shape of the potential. The kinetic term also acquires a more complex form which can significantly affect the scalar field dynamics [260]: zeros of the kinetic term can represent stationary values of ϕ. The ϕ field can be captured in the vicinity of such points in addition to minima of the potential.

In particular, as shown in [76], cosmological models with $\phi \to 0$ as $t \to \infty$, like those discussed in Sec. 10.3.1, become viable. As a result, we obtain a class of spatially flat cosmologies where both the observable scale factor $a(\tau)$ and the extra-dimensional one, $b(\tau)$ grow exponentially at late times, but $b(\tau)$ grows sufficiently slowly to leave variations of the effective gravitational constant G in the limits compatible with observations. Such models predict substantial changes in the physical laws in the remote future due to further growth of the extra dimensions.

To satisfy the tight experimental constraints on the temporal variation of the effective gravitational constant G, the parameters of the theory (10.75) should satisfy the condition [76]

$$12\left[\frac{d_1+2}{d_1} + \frac{8(c_1+c_2)}{d_1^2}\right] \gtrsim 100. \tag{10.81}$$

Thus, if we leave only $F(R)$ in the initial theory, it is impossible to satisfy the constraint (10.81): to do so, it is necessary to invoke the Ricci tensor squared or/and the Kretschmann scalar with the initial constants c_1 and c_2 such that

$$c_1 + c_2 \gtrsim d_1^2. \tag{10.82}$$

Under the condition (10.82), the model is potentially viable since it combines a de Sitter expansion of the observable Universe with sufficiently slow variations of the gravitational constant.

10.2.4 The second generalization: Several extra factor spaces

Let us consider nonlinear multidimensional gravity in more general spaces, admitting a reduction to smaller dimensions, namely, in a D-dimensional space-time $\mathbb{M}$ with the structure

$$\mathbb{M} = \mathbb{M}_0 \times \mathbb{M}_1 \times \cdots \times \mathbb{M}_n, \tag{10.83}$$

where $\dim \mathbb{M}_i = d_i$, and the metric

$$ds_D^2 = g_{ab}(x)dx^a dx^b + \sum_{i=1}^{n} \mathrm{e}^{2\beta_i(x)} g^{(i)}, \tag{10.84}$$

where (x) denotes the dependence on the first d_0 coordinates x^a; $g_{ab} = g_{ab}(x)$ is the metric in $\mathbb{M}_0$, and $g^{(i)}$ are x-independent d_i-dimensional metrics of the factor spaces $\mathbb{M}_i$, $i = \overline{1,n}$.

The initial action is taken in the form

$$S = \frac{1}{2} m_D^{D-2} \int \sqrt{^D g}[F(R) + c_1 R_{AB}R^{AB} + c_2\mathcal{K} + L_m], \tag{10.85}$$

where $F(R)$ is an arbitrary function of the scalar curvature R of the space $\mathbb{M}$; c_1, c_2 are constants; R_{AB} and $\mathcal{K} = R_{ABCD}R^{ABCD}$ are the Ricci tensor and the Kretschmann scalar of the space $\mathbb{M}$, respectively; and L_m is the matter Lagrangian. Capital Latin indices encompass all D coordinates, small Latin ones $(a, b, \ldots)$ the coordinates of the factor space $\mathbb{M}_0$, and a_i, $b_i, \ldots$ the coordinates of the spaces $\mathbb{M}_i$.

We will suppose that the "Kaluza–Klein" (KK) factor spaces $\mathbb{M}_i$ are compact and small in size, to exclude their observability by modern experimental tools.

The D-dimensional Riemann tensor has the nonzero components

$$
\begin{aligned}
R^{ab}{}_{cd} &= \overline{R}^{ab}{}_{cd},\\
R^{aa_i}{}_{bb_i} &= \delta^{a_i}_{b_i}\, B^a_{b\,(i)}, \quad B^a_{b\,(i)} := \mathrm{e}^{-\beta_i}\nabla_b(\mathrm{e}^{\beta_i}\beta_i^{,a}),\\
R^{a_ib_i}{}_{c_id_i} &= \mathrm{e}^{-2\beta_i}\overline{R}^{a_ib_i}{}_{c_id_i} + \delta^{a_ib_i}{}_{c_id_i}\beta_{i,a}\beta_i^{,a},\\
R^{a_ib_k}_{c_id_k} &= \delta^{a_i}_{c_i}\delta^{b_k}_{d_k}\beta_{i,a}\beta_k^{,a}, \quad i \neq k.
\end{aligned}
\tag{10.86}
$$

The overbar marks quantities obtained from the factor space metrics g_{ab} and $g^{(i)}$ taken separately, $\beta_{,a} \equiv \partial_a\beta$; it is denoted, as before, $\delta^{ab}{}_{cd} \equiv \delta^a_c\delta^b_d - \delta^a_d\delta^b_c$ and similarly for other kinds of indices.

The nonzero components of the Ricci tensor and the scalar curvature are

$$
\begin{aligned}
R^b_a &= \overline{R}^b_a + \sum_i d_i\, B^b_{a\,(i)},\\
R^{b_i}_{a_i} &= \mathrm{e}^{-2\beta_i}\overline{R}^{b_i}_{a_i} + \delta^{b_i}_{a_i}[\Box\beta_i + \beta_{i,a}\sigma^{,a}], \quad \sigma := \sum_i d_i\beta_i,\\
R &= \overline{R}[g] + \sum_i \mathrm{e}^{-2\beta}\overline{R}_i + 2\Box\sigma + (\partial\sigma)^2 + \sum_i d_i(\partial\beta_i)^2,
\end{aligned}
\tag{10.87}
$$

where $\sum_i$ means $\sum_{i=1}^n$; $\sigma := \sum_i d_i\beta_i$; $(\partial\sigma)^2 \equiv \sigma_{,a}\sigma^{,a}$ and similarly for other functions; $\Box = g^{ab}\nabla_a\nabla_b$ is the d_0-dimensional d'Alembert operator; $\overline{R}[g]$ and $\overline{R}_i$ are the Ricci scalars corresponding to g_{ab} and $g^{(i)}$, respectively.

Let us assume that the factor spaces $\mathbb{M}_i$ are d_i-dimensional compact spaces of constant nonzero curvature $K_i = \pm 1$, i.e., spheres ($K_i = 1$) or compact d_i-dimensional hyperbolic spaces ($K_i = -1$) with a fixed curvature radius r_0, normalized to the D-dimensional analogue m_D of the Planck mass, i.e., $r_0 = 1/m_D$ (recall that we use the natural units $c = \hbar = 1$). Then we have

$$
\begin{aligned}
&\overline{R}^{a_ib_i}{}_{c_id_i} = K_i\, m_D^2\, \delta^{a_ib_i}{}_{c_id_i},\\
&\overline{R}^{b_i}_{a_i} = K_i\, m_D^2\,(d_i-1)\delta^{b_i}_{a_i}, \quad \overline{R}_i = K_i\, m_D^2\, d_i(d_i-1).
\end{aligned}
\tag{10.88}
$$

The scale factors $r_i(x) \equiv \mathrm{e}^{\beta_i}$ in (10.84) are dimensionless.

10.2.5 Slow-change approximation. Reduction to d_0 dimensions

Suppose that all quantities are slowly varying, i.e., consider each derivative ∂_a (including those in the definition of $\overline{R}$) as an expression containing a small parameter ε, and neglect all quantities of orders higher than $O(\varepsilon^2)$.

Then we have the following decompositions:

$$
\begin{aligned}
R &= \phi + \overline{R}[g] + f_1 \quad f_1 := 2\Box\sigma + (\partial\sigma)^2 + \sum_i d_i(\partial\beta_i)^2; \\
F(R) &= F(\phi) + F'(\phi)(\overline{R}[g] + f_1) + O(\varepsilon^4); \\
R_{AB}R^{AB} &= \sum_i \frac{1}{d_i}\phi_i^2 + 2\sum_i d_i\phi_i[\Box\beta_i + (\partial\beta_i, \partial\sigma)] + O(\varepsilon^4); \\
\mathcal{K} &= 2\sum_i \frac{\phi_i^2}{d_1(d_1-1)} + 4\sum_i d_i\phi_i(\partial\beta_i)^2 + O(\varepsilon^4),
\end{aligned} \tag{10.89}
$$

where

$$
\phi_i := K_i m_D^2 (d_i - 1)\,\mathrm{e}^{-2\beta_i}, \quad \phi := \sum_i d_i\phi_i. \tag{10.90}
$$

The symbol $(\partial\alpha, \partial\beta)$ means $g^{ab}\alpha_{,a}\beta_{,b}$, and $F'(\phi) = dF/d\phi$.

As a result, neglecting $O(\varepsilon^2)$ and integrating out all $\mathbb{M}_i$, we obtain the action reduced to d_0 dimensions:

$$
\begin{aligned}
S &= \frac{1}{2}\mathcal{V}\, m_D^{d_0-2}\int \sqrt{g_0}\, d^{d_0}x\,\{\,\mathrm{e}^\sigma F'(\phi)\overline{R}_0 + K - 2V(\phi_i) + \mathrm{e}^\sigma L_m\}, \\
K &= F'(\phi)\,\mathrm{e}^\sigma f_1 + 2\,\mathrm{e}^\sigma \sum_i d_i\phi_i[c_1\Box\beta_i + c_1(\partial\beta_i, \partial\sigma) + 2c_2(\partial\beta_i)^2], \\
-2V(\phi_i) &= \mathrm{e}^\sigma\left[F(\phi) + \sum_i d_i\phi_i^2\left(c_1 + \frac{2c_2}{d_i - 1}\right)\right],
\end{aligned} \tag{10.91}
$$

where $g_0 = |\det(g_{\mu\nu})|$ and $\mathcal{V}$ is a product of volumes of n compact d_i-dimensional spaces $\mathbb{M}_i$ of unit curvature. The expression (10.91) is typical of a (multi)scalar-tensor theory (STT) of gravity in a Jordan frame.

Subtracting a full divergence, we get rid of second-order derivatives in (10.91), and the resulting kinetic term takes the form

$$
\begin{aligned}
K = {} & F'\,\mathrm{e}^\sigma\left[-(\partial\sigma)^2 + \sum_i d_i(\partial\beta_i)^2\right] - 2F''\,\mathrm{e}^\sigma(\partial\phi, \partial\sigma) \\
& + 4\,\mathrm{e}^\sigma(c_1 + c_2)\sum_i d_i\phi_i(\partial\beta_i)^2,
\end{aligned} \tag{10.92}
$$

where $F'' = d^2F/d\phi^2$.

Transition to the Einstein frame

For further analysis, it is helpful to pass on to the Einstein frame using the conformal mapping (10.60). The expression with the scalar curvature in (10.91) transforms as follows:

$$\sqrt{g_0}\, \mathrm{e}^{\sigma} \overline{R}_0 = \sqrt{g_0} f \overline{R}_0 = (\mathrm{sign}\, f)\sqrt{\widetilde{g}}\left[\widetilde{R} + \frac{d_0-1}{d_0-2}\,\frac{(\widetilde{\partial} f)^2}{f^2}\right] + \mathrm{div}, \qquad (10.93)$$

where the tilde marks quantities obtained from or with $\widetilde{g}_{\mu\nu}$ and div denotes a full divergence which does not contribute to the field equations. The action (10.91) acquires the form

$$S = \frac{1}{2}\mathcal{V}\, m_D^{d_0-2} \int \sqrt{\widetilde{g}}\, d^{d_0}x \{[\mathrm{sign}\, F'(\phi)][\widetilde{R} + K_E] - 2V_E(\phi_i) + L_{m(E)}\}, \qquad (10.94)$$

with the kinetic and potential terms

$$K_E = \frac{1}{d_0-2}\left(\partial\sigma + \frac{F''}{F'}\partial\phi\right)^2 + \left(\frac{F''}{F'}\right)^2 (\partial\phi)^2 + \sum_i \left[1 + \frac{4}{F'}(c_1+c_2)\phi_i\right](\partial\beta_i)^2, \qquad (10.95)$$

$$-2V_E(\phi_i) = \mathrm{e}^{-2\sigma/(d_0-2)} |F'|^{-d_0/(d_0-2)} \left[F(\phi) + \sum_i d_i\phi_i^2 \left(c_1 + \frac{2c_2}{d_i-1}\right)\right], \qquad (10.96)$$

where the tildes are omitted though the metric $\widetilde{g}_{\mu\nu}$ is used, and the indices are raised and lowered with $\widetilde{g}_{\mu\nu}$. The matter Lagrangian is

$$L_{m(E)} = \mathrm{e}^{-2\sigma/(d_0-2)} |F'|^{-d_0/(d_0-2)} L_m. \qquad (10.97)$$

The quantities β_i and σ are expressed in terms of the n fields ϕ_i, whose number coincides with the number of factor spaces.

10.3 Extra dimensions and low-energy physics

The introduction of extra dimensions is justified by the fact that their existence explains a great number of observed effects in the most economic way. In what follows we will present different applications of the general results of mathematical nature described above. In all our constructions we do not use any fields other than gravity. It should be noted that, although the idea of extra dimensions is often connected and even identified with string theory, in our consideration the idea of strings is not employed. We should also warn the reader that we are not building any final and complete theory here. Our goal is to demonstrate the wealth of opportunities that follow from the idea of extra dimensions. We therefore, at the present stage, do not fix a particular form of the initial Lagrangian and the number of extra dimensions. This allows us to explain one or another phenomenon in the most economic way.

It is assumed that a D-dimensional space-time is born as a result of quantum fluctuations from space-time foam. Thus there can appear spaces with different topologies, geometry and dimensionality. A calculation of probabilities of such processes does not seem productive at the present stage, and we simply use the assumption that these probabilities are not zero.

10.3.1 Self-stabilization of an extra space

One of the shortcomings of the idea of extra dimensions is the difficulty in explaining why they are stable. The reasons why the size of extra dimensions remains small and finite instead of collapsing to the Planck value or expanding to visible values remain vague. We will show how this problem can be solved in the framework of our approach.

In what follows we consider pure gravity ($L_m = 0$) and use the units $m_D = 1$, thus dealing with dimensionless quantities.

As the first example we consider the action (10.53) with the function

$$F(\phi) = \phi + c\phi^2 - 2\Lambda, \quad c,\ \Lambda = \text{const.} \tag{10.98}$$

Then (10.66) and (10.67) give the effective potential (10.80)

$$V^{(2)}_{\text{Ein}}(\phi) = -\frac{\text{sign}(1+2c\phi)}{[d_1(d_1-1)]^{d_1/2}}|\phi|^{d_1/2}\frac{c\phi^2+\phi-2\Lambda}{(1+2c\phi)^2}, \tag{10.99}$$

and the coefficient K_{Ein} of the kinetic term

$$K^{(2)}_{\text{Ein}}(\phi) = \frac{1}{(1+2c\phi)^2\phi^2} + \left[c^2\phi^2(d_1^2-2d_1+12) + d_1^2 c\phi + \frac{1}{4}d_1(d_1+2)\right]. \tag{10.100}$$

The potential is plotted in Figs. 10.1 and 10.2.

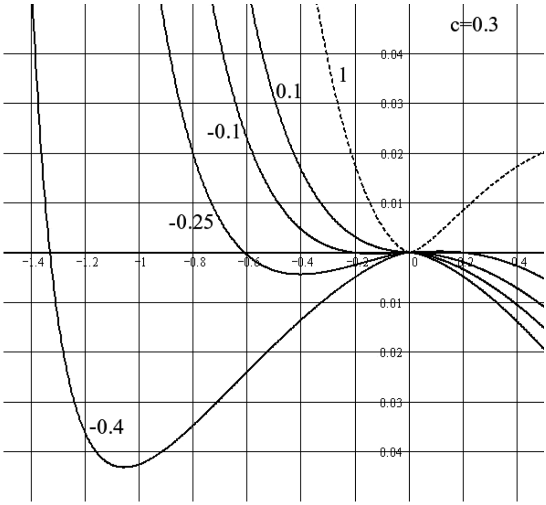


Figure 10.1 The potential $V_{\rm Ein}(\phi)$ for $d_1 = 3$, $c = 0.3$ in the range $\phi > -5/3$, where $F'(R) > 0$, for a few values of Λ (indicated near the curves). The dashed line ($\Lambda = 1$) shows a minimum at $\phi = 0$. The values of ϕ are expressed in the units $m_D = 1$.

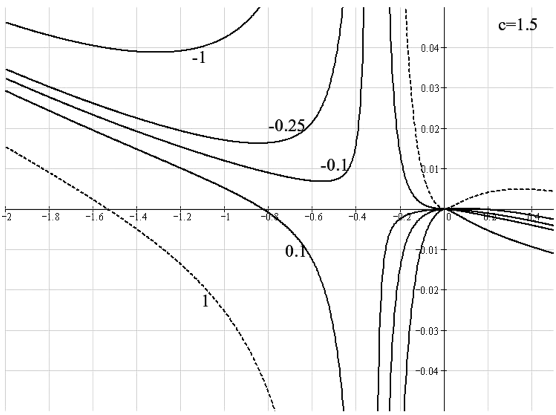


Figure 10.2 The same as in Fig. 10.1, but with $c = 1.5$. Minima of $V(\phi)$ are located in the range $\phi < -1/3$, where $F' < 0$.

The existence of minima of the potential indicates possible stabilization of the size of extra dimensions. Indeed, if a field is in a minimum of the effective potential, there is a solution of the form $\phi(t) = \text{const}$. And since the size of the extra space is $L = b \sim 1/\sqrt{\phi}$, it means that this size is fixed.

There is one more minimum of the potential $V_{\rm Ein}$ with $F' > 0$, existing in the range $\Lambda > 0$ and located at the point $\phi = 0$. The asymptotic $\phi \to 0$ corresponds to growing rather than stabilized extra dimensions: $b = {\rm e}^{\beta} \sim 1/\sqrt{|\phi|} \to \infty$. A model with such an asymptotic growth at late times may still be of interest if the growth is sufficiently slow and the size b does not reach detectable values by now.

Potentials with $\Lambda < 0$ (Fig 10.2) show minima with $V_{\min} > 0$, so that $\Lambda_{\rm eff} > 0$, which should lead to cosmological models with de Sitter external space and stable extra dimensions. Such models again correspond to $\phi < 0$, i.e., hyperbolic extra dimensions. By fine tuning of the initial constants c and Λ we can obtain the present-day values of the cosmological parameters. As in Fig. 10.1, we also find a minimum at $\phi = 0$, whose properties have already been discussed.

Thus the behavior of the system is drastically different in different ranges of ϕ and depends on the numerical values of the initial parameters. We also confirm that gravity alone can stabilize the size of extra dimensions, without the need for introducing other fields with specific forms of potentials. Positive values of $\Lambda_{\rm eff}$ (vacuum energy density) were found in the range where $F' < 0$. As will be seen below, it is a feature of quadratic gravity only.

10.3.2 On the influence of the number of extra dimensions on low-energy physics

We have discussed some low-energy theories corresponding to special cases of the metric (10.84) at different choices of the initial action. It is not surprising that some values of the parameters are suitable for describing some properties of our Universe. More opportunities emerge if we vary the structure of extra dimensions, including the number of extra factor spaces, their dimension and curvature. In such cases it is possible to obtain a set of substantially different low-energy theories *even with a fixed initial Lagrangian*.

Let us recall once again the well-known shortcoming of chaotic inflation in its simplest, quadratic form. According to the observable temperature fluctuations of the CMB, the inflaton mass should be of the order $10^{-6}\,M_{\rm Pl}$. Its smallness requires an explanation which so far does not exist. Consider the effective potential (10.96), created by the initial action (10.64) and (10.98). Its shape is presented in Fig. 10.3 for some values of the number of extra dimensions d. It is evident that the values of d can be chosen in

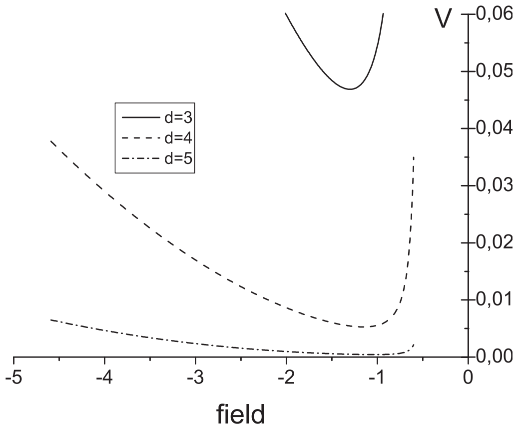


Figure 10.3 Effective potential $V(\phi)$ for the theory (10.64) with $F(R)$ of the form (10.98) for different numbers d of extra dimensions. The parameter values: $c=1,\ \Lambda=-0.4$.

such a way as to obtain the required inflaton mass. Numerical estimates lead to the following results: the second-order derivative of the potential at its minimum (which determines the inflaton mass) is $\sim 0.2m_D$ for $d=3$, $\sim 1.5\cdot 10^{-3}m_D$ for $d=5$ and $\sim 0.8\cdot 0^{-5}m_D$ for $d=7$. The initial Lagrangian does not contain small parameters. Nevertheless the required value of the inflaton mass emerges at a classical level at a suitable choice of the parameter d.

The strong influence of the number d on the form of the low-energy Lagrangian is not surprising because the potential (10.96) contains the factor $\sim d^{-d}$. Thus, if the stationary state is $\phi=\phi_0$, and the dimensionless parameters $|\phi_0|m_D^{-2}$ and $F'(\phi_0)$ are of the order of unity, then the effective cosmological constant $\Lambda_{\text{eff}}=V_{\text{Ein}}(\phi_0)$ is connected with the function $F(\phi_0)$ (which can be close to m_D^2) as follows:

$$\Lambda_{\text{eff}}/F(\phi_0)\sim[d(d-1)]^{-d/2}. \tag{10.101}$$

It is of interest that $d^{-d}\approx 10^{-123}$ for $d=67$. It means that space-time fluctuations creating a (67+4)-dimensional space can cause the creation of space with the observable vacuum energy density $10^{-123}\,m_4$. The extreme smallness of Λ_{eff} is thus related to the number of extra dimensions of a space randomly created by quantum fluctuations at the Planck level.

Figure 10.4 shows another example of d-dependence on the effective potential shape. It is evident that the very existence of a minimum of

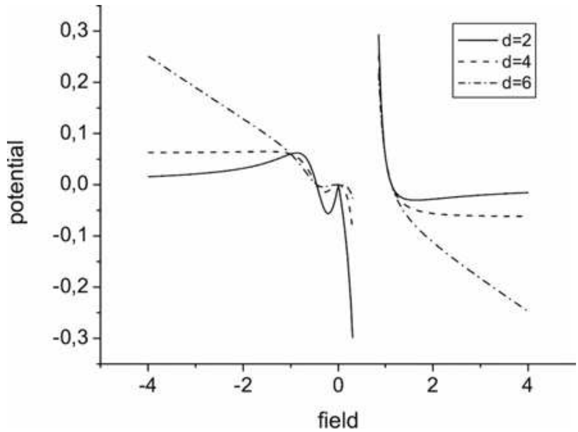


Figure 10.4 Potentials at different numbers d of extra dimensions, obtained from the Lagrangian (10.64) with $F(R) = R + cR^2 + w_1R^3 + w_2R^4 - 2\Lambda$. The parameters: $c = 0,\ w_1 = 0,\ w_2 = -1,\ \Lambda = -0.25$. The curves are adjusted to a single scale.

the potential, indicating a possible stable configuration, is d-dependent. Thus, if the Universe was born with a single extra factor space of negative curvature, then $\phi < 0$, see Fig. 10.4. It is clear that with the initial field value $\phi < -1$ its mean value in such a Universe tends to infinity if $d = 2$, to a constant if $d = 4$ and to zero if $d = 6$. If the Universe was born with $-1 < \phi < 0$, then the mean value of the field is localized at a minimum of the potential, and the size of the extra space remains fixed.

It is known that the existence of local minima of the potential corresponds to a possible false vacuum decay . As an example, Fig. 10.5 shows the effective potential (10.80), emerging under the condition that the extra space has the topology of a sphere. It is evident that the field located at a metastable minimum will eventually pass on to the basic state with $\phi = 0$. It means that after the phase transition the size of the extra space will grow infinitely. Nevertheless, the Universe existing in a metastable state can survive for quite a long time if the probability of tunneling to the lower minimum is small.

10.3.3 Extra dimensions and inflation

A single extra factor space

Now, our program is as follows:

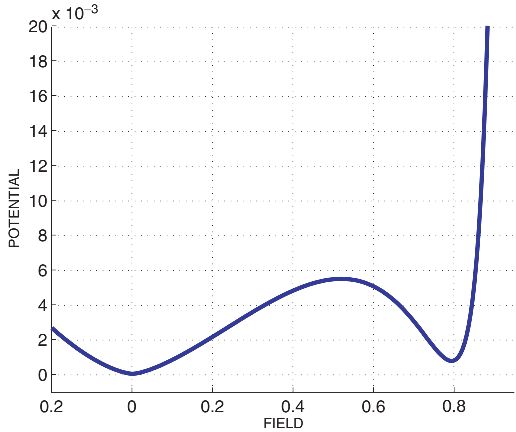


Figure 10.5 The effective potential (10.132). The parameter values: $c = -0.5$, $\Lambda = 0.2$, $c' = -0.626$, and $d = 2$.

1. Choose the parameters of the original action (10.85) to obtain a behavior of the potential (10.96) providing primordial inflation.
2. Additionally vary the parameters to satisfy the inflationary conditions conforming to observations.
3. Try to describe the modern acceleration stage, providing the ratio of the effective cosmological constant to the Planck density $\Lambda_{\text{eff}}/m_4^4$ of the order 10^{-123}.

We begin with the case of one factor space. Then Eqs. (10.95) and (10.96) simplify to give

$$S = \frac{\mathcal{V}}{2}\int d^4x\,\sqrt{g}\,(\operatorname{sign} F')L,$$

$$L = \widetilde{R}_4 + K_E^{(1)}(\phi)(\partial\phi)^2 - 2V_E^{(1)}(\phi), \tag{10.102}$$

$$K_E^{(1)}(\phi) = \frac{1}{4\phi^2}\left[6\phi^2\left(\frac{F''}{F'}\right)^2 - 2d_1\phi\frac{F''}{F'} + \frac{1}{2}d_1(d_1+2)\right]\frac{c_1+c_2}{F'\phi}, \tag{10.103}$$

$$V_E^{(1)}(\phi) = -\frac{\operatorname{sign} F'}{2F'^2}\left[\frac{|\phi|}{d_1(d_1-1)}\right]^{d_1/2}\left[F(\phi) + c_V\frac{\phi^2}{d_1}\right],$$

$$c_V \equiv c_1 + \frac{2c_2}{d_1-1}. \tag{10.104}$$

Here we take

$$F = F(\phi) = \phi + c\phi^2 - 2\Lambda, \quad c, \ \Lambda = \text{const}, \tag{10.105}$$

and $\phi = d_1\phi_1$.

In (10.102)–(10.104) we have actually changed the sign of the Lagrangian in case $F' < 0$; to preserve the attractive nature of gravity for ordinary matter, the matter Lagrangian density should appear with an unusual sign from the beginning. As a result, the sign of the whole action of gravity and matter will be unusual, without any effect on the equations of motion. As was discussed above, one can show that quantum transitions are then unaffected as well.

The presence of the parameters c_1 and c_2 adds freedom in choosing the shape of the potential. The kinetic term also has a complex form which can substantially affect the field dynamics. An analysis of kinetic terms like (10.103) of variable sign can lead to interesting possibilities, and we hope to return to this point in our future work.

Let us employ the fact that chaotic inflation with a quadratic potential and the inflaton mass $m_\varphi \approx 10^{-6} m_4$ well conforms to the observational data. Therefore our task is simplified and reduced to finding such parameters c, c_1 and c_2 that the potential (10.104) near its minimum is approximated by a quadratic function with the above inflaton mass. It turns out to be possible with the following parameter values:

$$d_1 = 4; \quad c = 2.5 \cdot 10^4; \quad c_1 + c_2 = 0.6;$$
$$c_V = -2.48, \quad \Lambda = 0.2. \tag{10.106}$$

With these parameter values, all basic requirements to inflation are satisfied. Thus, the duration of the inflationary period exceeds 60 e-folds, the temperature fluctuations are $\sim 6 \cdot 10^{-5}$, and the spectral index is $n_s = 0.943$, within observational bounds, $n_s = 0.958 \pm 0.016$. Thus a single factor space is sufficient for obtaining a fairly good inflationary scenario.

Since the constant c has actually the dimension of length squared, it is $\sqrt{c} \sim 100$ that should be compared with the Planck length. So this model does not contain unnaturally large or small parameters.

A serious shortcoming of this model is that it is unable to solve the problem of modern acceleration, including the smallness of dark energy density. Indeed, it is easy to prove that slight variations of the parameters c, c_1 and c_2 could give rise to an arbitrarily small potential value at the minimum. However, though the values of these parameters are quite moderate, they need to be extremely "fine-tuned" to fit the modern value of vacuum energy density. An attempt to solve this problem in a slightly more complex model is undertaken in the next section.

10.3.4 Two factor spaces: Inflation and modern acceleration

Inflation

Additional opportunities emerge if the extra space is a product of two factor spaces, $\mathbb{M}_{d_1} \times \mathbb{M}_{d_2}$ of dimensions d_1 and d_2. For further analysis, let us make the situation more specific by putting $K_1 = K_2$, $d_1 = d_2$ and choosing the function

$$F(R) = R^2. \tag{10.107}$$

(Note that one of the coefficients in the initial Lagrangian can be chosen at will, e.g., equal to unity, without affecting the field equations; it simply specifies the scale for other coefficients.)

Figure 10.6 presents the potential of the effective scalar fields for this model with the following choice of the parameter values:

$$d_1 = d_2 = 5, \quad c_V = -10.001, \quad c_1 + c_2 = 1.25 \cdot 10^3. \tag{10.108}$$

All further numerical estimates will be obtained with these values. As follows from the above, at low energies (as compared to the Planck scale m_D) this model is equivalent to Einstein gravity with two scalar fields. In full similarity with the previous section, the constants c_1 and c_2 have actually the dimension of length squared, and their square roots are not unnaturally large or small.

Note that, with the c_V value chosen, a positive potential V (hence a positive effective cosmological constant) is obtained with $K_1 = 1$, i.e., a spherical extra factor spaces. For other values of c_V, e.g., $c_V > 0$ we would need hyperbolic factor spaces.

The inflationary period is characterized by moving down one of the steep slopes of the valley. The inflaton mass squared is proportional to the second-order derivative of the potential in the direction perpendicular to the valley (its bottom is located at $\phi_1 = \phi_2 = \phi_0$). It is this direction in which the field moves during inflation and oscillates during reheating at the post-inflationary stage. The specific value of ϕ_0 depends on the initial value of the inflaton field at which the classical universe was born.

Figure 10.7 shows the dependence of the effective inflaton mass on the parameter ϕ_0. In the framework of chaotic inflation, universes are created

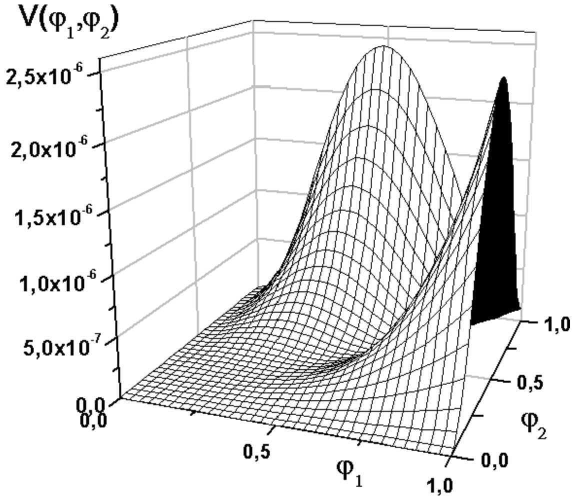


Figure 10.6 Potential of the effective scalar fields for the model (10.85) and (10.107) with the parameter values given in (10.108).

with different inflaton values under the horizon, leading to different values of ϕ_0 and hence different inflaton masses. This is how this model solves the problem of smallness of the inflaton mass in Planck units. It is a point of importance: instead of introducing a small parameter (the inflaton mass) and wondering why it is small, we have obtained a self-fitting mechanism. We live in a particular Universe, among many, with the appropriate inflaton mass.

Post-inflationary particle production is a result of oscillations in the direction across the valley. The conditions suitable for our Universe correspond to the value $\phi_0 \simeq 0.5$. It is just such a value that, according to Fig. 10.7, the inflaton mass, related to the second-order derivative of the potential in the direction across the valley, is $\sim 10^{-6}$–10^{13} GeV, which satisfactorily explains the observational data on CMB temperature fluctuations.

It is of interest to what extent the values of c_1 and c_2 (or, more conveniently, their combinations c_V and $c_K = c_1+c_2$) in (10.108) are fine-tuned. An inspection shows that with c_V in the range $(-10.2, -10)$ the potential provides all three necessary stages of evolution: inflation, reheating and the present expansion, in agreement with the observational data under proper initial conditions. Larger deviations destroy the valley of the potential surface thus drastically changing the whole picture.

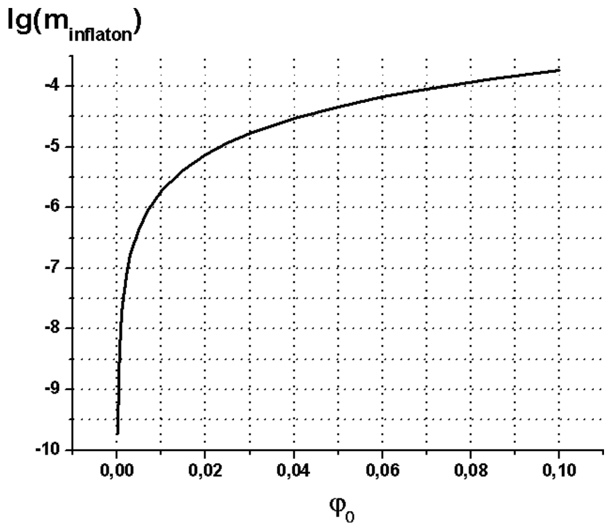


Figure 10.7 ϕ_0 dependence of the effective inflaton mass (Planck units).

The admissible range of c_K is wider: its value may vary by an order of magnitude with respect to the one given in (10.108). Within this area, the predictions are actually the same, within uncertainties in the observational data.

Matter-dominated stage

The inflationary stage ends with rapid field oscillations across the valley in Fig. 10.6, on whose bottom, by our assumptions, $\phi_1 = \phi_2 = \phi/(2d_1)$. These oscillations are accompanied by effective particle production in full agreement with the standard version of chaotic inflation with a quadratic potential. In the model under discussion, the energy density of the produced particles makes the material content of the Universe and affects not only the cosmological expansion rate but also the scalar field dynamics. The latter now corresponds to slowly rolling down along the bottom of the potential valley.

We assume a spatially flat cosmology in 4 dimensions, with the Einstein-frame metric $d\widetilde{s}_4^2 = dt^2 - a^2(t)d\vec{x}^2$. So, with the choice (10.107), the action (10.94) leads to the Lagrangian

$$L_E = R_4 + K(\phi)(d\phi)^2 - 2V(\phi), \qquad (10.109)$$

and

$$K(\phi)(\partial\phi)^2 = K_0(\partial\phi)^2/\phi^2 = 4K_0(\partial\beta)^2,$$
$$2K_0 = d_1^2 - d_1 + 3 + 4(c_1 + c_2), \tag{10.110}$$
$$V(\phi) = V_0|\phi|^{d_1} = V_1\,\mathrm{e}^{-2d_1\beta},$$
$$V_1 = -\frac{K_1}{8}\left(1 + \frac{c_V}{2d_1}\right), \quad V_0 = V_1[2d_1(d_1-1)]^{-d_1}, \tag{10.111}$$

where $K_1 = \operatorname{sign}\phi = \operatorname{sign} F'(\phi)$ and $\beta(t) = \beta_1(t)$ is, as before, the logarithm of the extra-dimension scale factor (which is in the present case the same for all extra dimensions), such that $d\phi/\phi = -2d\beta$, and c_V has been defined in (10.104). One can also notice that a usual form of the Lagrangian with a scalar field Φ and a potential V_Φ is obtained if we substitute

$$2\sqrt{K_0}\beta = \sqrt{8\pi G}\Phi, \quad V_E = 8\pi G\, V_\Phi.$$

With (10.109), we can write two independent components of the Einstein-scalar equations for $\beta(t)$ and $a(t)$ as follows:

$$3H^2 = 2K_0\dot\beta^2 + V_1\,\mathrm{e}^{-2d_1\beta} + 8\pi\rho_m, \tag{10.112}$$
$$2K_0[\ddot\beta + 3H\dot\beta] = d_1V_1\,\mathrm{e}^{-2d_1\beta}, \tag{10.113}$$

where $H = \dot a/a$ is the Hubble parameter.

Let us begin with considering the matter-dominated stage, which is the longest. The subsequent dark energy (DE) dominated stage will be discussed in the next subsection. The following simplifying assumptions will be used: (i) we neglect the pressure of matter, treating it as dust from the very beginning ($t = t_1$) thus ignoring a radiation-dominated stage; (ii) we neglect a possible direct interaction between matter and the scalar field; (iii) we neglect the scalar field contribution to the dynamics of $a(t)$ at the matter dominated stage $t_1 < t < t_2 \simeq 10^{10}$ years and, vice versa, we neglect the contribution of matter at the DE dominated stage $t > t_2$.

So, neglecting the contribution of β in (10.112), we obtain for times $t_1 < t < t_2$, as in the usual Big Bang scenario,

$$H = 2/(3t) \quad \text{at } t_1 < t < t_2. \tag{10.114}$$

To solve Eq. (10.113) numerically, we take the following initial data corresponding to the end of the post-inflationary epoch:

$$\phi_1(t_1) = \phi_2(t_1) = \frac{\phi(t_1)}{2d_1} = 0.05; \quad \frac{d\phi}{dt}(t_1) = 0$$
$$\Rightarrow \mathrm{e}^{\beta(t_1)} = 4\sqrt{5}, \quad \dot\beta(t_1) = 0. \tag{10.115}$$

The initial time t_1 is chosen to be $t_1 = 9\cdot 10^9$ for definiteness.

Numerical solution of Eq. (10.113) then gives the following value of β at $t = t_2$:

$$\mathrm{e}^{\beta(t_2)} \simeq 5.5{\cdot}10^{11} \Rightarrow \phi(t_2) \simeq 1.3{\cdot}10^{-22}. \tag{10.116}$$

This value of β will be used in analyzing its dynamics at the modern stage for which the equations simplify and can be solved analytically.

Modern stage

The modern epoch $t > t_2$ is DE dominated. In the present approach, the DE is represented by the scalar field ϕ (or equivalently β or $b = \mathrm{e}^{\beta}$) with the potential (10.110), and the Universe dynamics is described by Eqs. (10.112) and (10.113). In (10.112) we now neglect the matter contribution.

It is hard to solve this set of equations exactly. However, as the ϕ field decreases (which corresponds to a growing size of the extra dimensions) along with a decreasing value of the potential (related to the effective cosmological constant), at some stage it becomes possible to treat this process as secondary slow rolling, for which the field dynamics is sufficiently simple and may be described analytically. Indeed, let us suppose

$$|\ddot\beta| \ll 3(\dot a/a)\dot\beta, \quad K_0\dot\beta^2 \ll 3(\dot a/a)^2, \tag{10.117}$$

and drop the corresponding terms in Eqs. (10.112) and (10.113). Then we can express $\dot a/a$ from (10.112) and insert it to (10.113), getting

$$d_1\dot\beta\,\mathrm{e}^{d_1\beta} = B_0 := \frac{d_1^2\sqrt{V_1}}{2\sqrt{3}K_0}, \tag{10.118}$$

whence we find the evolution law for the extra-dimension scale factor

$$\mathrm{e}^{\beta} = [B_0(t - t_*)]^{1/d_1}, \tag{10.119}$$

where t_* is an integration constant ($t_* = t_2 - B_0^{-1}[b(t_2)]^{d_1}$). Substituting this result to (10.112), we find the evolution law for $a(t)$:

$$a(t) = a_*(t - t_*)^p, \quad p := 2K_0/d_1^2, \tag{10.120}$$

where a_* is an integration constant.

With the parameters (10.108), some relevant constants are

$$V_1 = 1.25{\cdot}10^{-4}, \quad 2K_0 = 5023,$$
$$p = \frac{5023}{25} \approx 201, \quad B_0 \approx 3.2{\cdot}10^{-5} \tag{10.121}$$

Eq. (10.119) with the initial value (10.116) gives the present size of the extra dimensions, at $t = t_0 = 13.7{\cdot}10^9$ yr:

$$b(t_0) \simeq 5.5 \cdot 10^{11} \approx 9{\cdot}10^{-22}\ \text{cm}, \tag{10.122}$$

well within the observational limits. From (10.120) we find the Hubble constant $H_0 = \dot{a}(t_0)/a(t_0)$ and the Hubble time $t_H = 1/H_0$:

$$H_0 \approx 1.25{\cdot}10^{-61}, \quad t_H \approx 8{\cdot}10^{60} \approx 13.8{\cdot}10^9\ \text{yr}, \tag{10.123}$$

in agreement with observations. The potential energy density V, coinciding with the DE density,

$$V_E(\phi(t_0)) \simeq 5.1{\cdot}10^{-123}, \tag{10.124}$$

also well agrees with observations.

One can notice that in our model with $d_1 = 5$ the function (10.119) grows extremely slowly. The present value in (10.122) differs from that in (10.116) only in the fifth decimal digit, so that the change is actually indistinguishable. The same is true for the DE density which thus behaves like a cosmological constant. The expansion law (10.120) with the exponent $p = 201$ is really almost exponential, i.e., de Sitter, and the DE equation-of-state factor $w = p_{\rm DE}/\rho_{\rm DE}$ is very close to minus unity. Indeed, in the DE epoch, $a(t) \sim t^{2/(3+3w)}$, hence

$$2/(3+3w) = 201 \Rightarrow w \approx -0.9967.$$

Lastly, one can verify that this solution fairly well satisfies the slow-rolling conditions (10.117), which hold as long as $p \gg 1$, or, in terms of the input parameters of the theory, if $c_1 + c_2 \gg d_1^2$.

It is of interest that models of gravity (10.85) where $F(R)$ contains a linear term do not lead to similar attractive results in the present approach.

Discussion

In the framework of pure curvature-nonlinear gravity with extra dimensions, it has been possible to describe (though only in a rough approximation) the entire evolution of the Universe beginning with an inflationary stage and ending with the modern accelerated stage with sufficiently small dark energy density. In doing so, it has been possible to avoid unnaturally small or large parameter values in the initial Lagrangian. The small values of the inflaton mass and especially that of DE density agreeing with observations have been obtained from a Lagrangian whose dimensionless parameters differ from unity by no more than two orders of magnitude.

Using a single extra factor space, it appears possible to explain the emergence of an inflaton, and choosing proper values of the parameter, it is possible to fulfil all requirements applicable to inflationary models and achieve an agreement with the observational data. However, to solve the problem of small DE density, it is necessary to invoke (at least) two extra factor spaces.

The inflationary stage with an appropriate inflaton mass is again well described. Indeed, field fluctuations create universes with different initial field values. The potential in Fig. 10.6 (i.e., at fixed values of the initial Lagrangian parameters) has different curvatures at different points of the valley, which correspond to different inflaton masses. We live in a universe created by a suitable field fluctuation whose evolution leads to the observable inflaton mass.

As to late-time evolution, it becomes possible to obtain in a natural way a small current value of the effective potential which plays the role of DE density (effective cosmological constant), $\Lambda_{\text{eff}} \sim 10^{-123}\ m_4^4$. The form of our late-time solution shows that the size of the extra dimensions is slowly growing in the modern epoch. In the remote future, this size, which is so far invisible to modern instruments, will grow to such values that will lead to drastic changes in the physical laws of our Universe. Let us stress, however, that such a model is only one particular possibility contained in our approach. There are other models where the extra dimensions are stable at late times [83] making the effective physical constants also invariable.

Our model with two factor spaces has the following advantages:

(a) Its low-energy limit represents the Hilbert–Einstein action with appropriate accuracy.

(b) It describes inflaton with an inflation mass that agrees with observations.

(c) The size of the extra dimensions $b(t)$ never exceeded the experimental threshold $\sim 10^{-17}$ cm (though it should in the remote future).

(d) At the modern stage, the scalar field density (actually, the potential $V(\phi)$ in proper units) describes the modern DE density $\sim 10^{-123} m_4^4$;

(e) The DE equation-of-state parameter w satisfies the observational constraint $w < -0.8$.

This model has a somewhat unusual total dimension $D = d_0 + 2d_1 = 14$. With such a choice, it is clear that we do not use the ideas of string theory in this study, which makes our model less restrictive in the choice of the dimensionality. We believe that other choices of parameter sets, including dimensionality, can also lead to good potentials in the low-energy limit, and this can be a subject of future work. However, the particular values $d_1 + d_2 = 3 + 3$, leading to $D = 10$ (the "string" dimension), are probably unsuitable in our case since then $c_V = c_K = c_1 + c_2$ (see the notations above), which leaves one independent parameter instead of two thus substantially restricting the choice of effective potentials.

Since we have been working in the Einstein conformal frame, the problem of varying physical constants (above all, the effective Newtonian constant of gravity $G_{\rm eff}$) did not emerge. One should note that even remaining in the Einstein frame, we could assume $m_D \neq m_4$, which would affect the estimated boundary between the classical and quantum worlds. In a more general framework, interpreting another conformal frame (possibly but not necessarily the original Jordan frame) as the observational one, we obtain a dependence of the constant $G_{\rm eff}$ (hence the current Planck mass $m_4 = G_{\rm eff}^{-1/2}$) on the size of extra dimensions, which in general can not only be time-dependent but also vary from point to point in space. In the cosmological context, models with variable $G_{\rm eff}$ should not only satisfy the observational bounds on the variation rate $\dot{G}_{\rm eff}/G_{\rm eff}$ ($\lesssim 10^{-13}$) but also take into account the effect of $G(t)$ on stellar evolution and processes in the early Universe. (Therefore, models with self-stabilizing extra dimensions like those discussed in [77, 83] can be more attractive.) In still more general models of this sort even the Planck constant $\hbar$ can be variable. A discussion of these problems can be found, e.g., in [147, 358, 411].

10.3.5 Rapid particle creation in the post-inflationary period

Let us address one more problem that can be successfully solved in the class of theories under consideration.

According to [137], rapid oscillations of the inflaton field immediately after inflation create the particles which form the matter content of the modern Universe. It is known [243, 372] that the particle and entropy production mechanism as well as heating of the Universe can be ineffective if the coupling constant of the inflaton and matter fields is too small. On the other hand, a strong coupling leads to significant quantum corrections of the inflaton potential parameters, which casts doubt on their smallness that follows from the observational data on CMB temperature fluctuations. This problem is solved, for instance, by hybrid inflation which includes one more hypothetic scalar field (see Chapter 8). During inflation, the energy density changes slowly until the classical field value reaches a bifurcation point, after which the field begins to rapidly oscillate near a minimum of the potential, producing the required particle number. Unfortunately, this model also leads to black hole over-production, which is a serious shortcoming [354]. Another promising mechanism of effective particle production (parametric resonance) is described in [243]. One more mechanism, connected with quantum production of supermassive (GUT scale) particles, sufficient for explaining the observed wealth of matter, has been considered by Grib and Pavlov [179–182]

Nonlinear multidimensional gravity provides one more opportunity.

Consider the potential and kinetic terms of the effective Lagrangian (10.64) with the function $F(R)$ of the form

$$F(R) = R + cR^2 - 2\Lambda. \tag{10.125}$$

These functions are plotted in Fig. 10.8. The nontrivial ϕ dependence of the kinetic term strongly affects the scalar field dynamics. The effect is especially evident if a minimum of the kinetic term coefficient coincides with a minimum of the potential, as in Fig. 10.8.

For illustration, consider a simplified scalar field model with the potential and a coefficient in the kinetic terms of the form

$$\begin{aligned} V(\phi) &= \frac{1}{2}m^2\phi^2 \\ K(\phi) &= K_1 \cdot (\phi - \phi_{\min})^2 + K_{\min}, \quad K_1,\ K_{\min} > 0. \end{aligned} \tag{10.126}$$

(The function $K(\phi)$ is positive and can be reduced to unity by a field redefinition $\phi \to \tilde{\phi}$, but this equivalent formulation is more convenient in the present case.)

A numerical solution of the corresponding classical equations for a spatially flat universe,

$$\begin{aligned} H^2 &= \frac{\varkappa^2}{3}\left[\frac{1}{2}K(\phi)\dot{\phi}^2 + V(\phi)\right], \\ &K(\phi)(\ddot{\phi} + 3H\dot{\phi}) + \frac{1}{2}K_\phi(\phi)\dot{\phi}^2 + V_\phi(\phi) = 0, \end{aligned} \tag{10.127}$$

(the subscript ϕ denotes $d/d\phi$, and H is the Hubble parameter) is presented in Fig. 10.9 for the quadratic potential (10.126) and two different

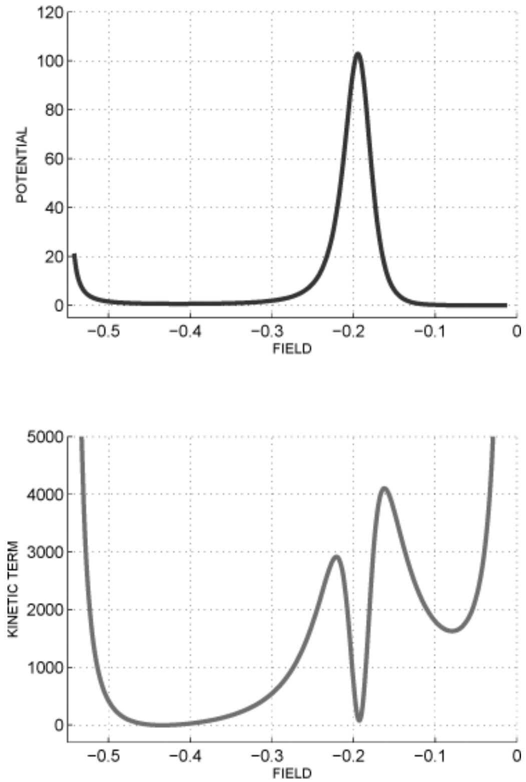


Figure 10.8 The potential and kinetic terms of the effective 4D theory (10.53) with the function $F(R)$ of the form (10.125) with the parameters $d = 3, c = 6.0, w_1\Lambda = 2$.

kinetic terms, with $K = 1/2$ and the one given in (10.126). The parameter values are given in the figure caption. Each curve describes the inflaton behaviour before and after the end of inflation. In the nonstandard kinetic term, $K(\phi) = 1/2$ at the end of the inflationary stage (the first point of curves' intersection in Fig. 10.9). It is clear that in the model with a variable kinetic term inflation begins at smaller energies as compared to the conventional case.

Let us also note the difference in the oscillation frequency of the inflaton field. The number of created particles is proportional to this frequency, see the review [139] and an additional discussion in [234]. This conclusion can also be grounded as follows.

After the end of inflation, the inflaton oscillation amplitude is small compared with the Planck scale, and the coefficient $K_{\text{eff}} \sim K_{\min}$ in the

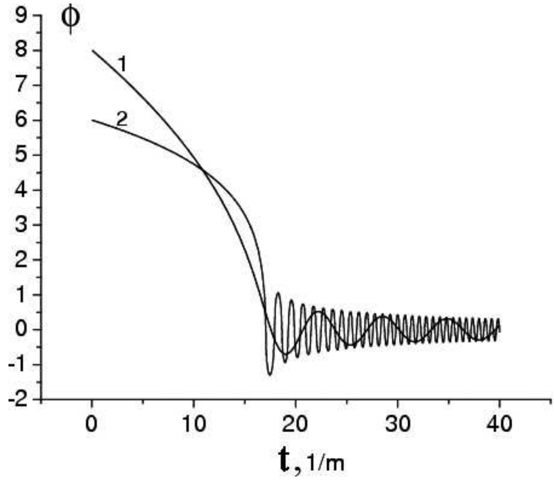


Figure 10.9 The inflaton behaviour at the end of inflation for two models: 1. The standard kinetic term, $K = 1/2$, 2. The kinetic term of the form (10.126). The parameter values $K_0 = 0.1, \phi_{\rm min} = 0.1, K_{\rm min} = 0.005$ are given in Planck units, and time is measured in inverse inflaton field masses.

effective kinetic term is small due to the choice of the parameters. The effective Lagrangian containing an interaction of the inflaton with a certain scalar field χ

$$L_{\rm eff} \simeq \frac{1}{2} K_{\rm min} \dot{\phi}^2 + g\phi\chi\chi, \tag{10.128}$$

can be transformed to the standard form by redefining the inflaton field:

$$L_{\rm eff} \simeq \frac{1}{2} \dot{\phi}^2 + \frac{g}{\sqrt{K_{\rm min}}} \phi\chi\chi. \tag{10.129}$$

A small value of $K_{\rm min}$ increases the coupling constant (by orders of magnitude for the chosen parameter values) and consequently leads to intense particle creation. In this way one overcomes the difficulty discussed above: a weak interaction constant at the inflationary stage does not contradict a rapid particle production immediately after the end of inflation.

The field behavior near $K = 0$

The function $K(\phi)$, the coefficient of the kinetic term in the scalar field Lagrangian, can in general possess both zeros and poles, and their neighborhood is of special interest. Let us discuss the possible state of the system near a zero value of $K(\phi)$ [260]. If it is a simple zero, then the kinetic term changes its sign when crossing it. The case $K(\phi) = K_{\rm Ein} > 0$ corresponds to a normal scalar field while the case $K(\phi) < 0$ corresponds to a phantom field with negative kinetic energy.

In general, $F(R)$ theories lead only to positive values of $K(\phi)$. Indeed, it is easy to see that the expression in square brackets in Eq. (10.66) is always positive, with a minimum equal to $d(d+3)/3$.

In more general theories containing other curvature invariants, this is not the case. Let us consider as an example the simplest nonlinear theory with

$$F(R) = R + cR^2 - 2\Lambda, \tag{10.130}$$

including second-order nonlinear terms: the Ricci tensor squared $R_{AB}R^{AB}$ and the Kretschmann scalar $\mathcal{K} = R_{ABCD}R^{ABCD}$, so that

$$S = \int d^D x \sqrt{{}^D g}[F(R) + c_1 R_{AB}R^{AB} + c_2 \mathcal{K}]. \tag{10.131}$$

Calculations in the model (10.131) with the function (10.130) and the metric (10.49), similar to those made above, lead to an effective scalar-tensor theory (10.64) with kinetic (10.79) and potential (10.80) terms

$$V_{\rm Ein}(\phi) = -\,\mathrm{sign}(1+2c\phi)[d(d-1)]^{-d/2}\cdot|\phi|^{d/2}\frac{c'\phi^2+\phi-2\Lambda}{(1+2c\phi)^2},$$

$$c' = c + \frac{c_1}{d} + \frac{2c_2}{d(d-1)}, \tag{10.132}$$

$$K_{\rm Ein}(\phi) = \frac{1}{\phi^2(1+2c\phi)^2}\left[c^2\phi^2(d^2-2d+12) + d^2c\phi + \frac{1}{4}d(d+2)\right] + \frac{c_1+c_2}{2\phi(1+2c\phi)}. \tag{10.133}$$

The dashed lines A and B in Fig. 10.10 mark zeros of the kinetic term for the quadratic gravity (10.130). If a field fluctuation has emerged to the left of point A, the field tends to a minimum of potential at the point $\phi = 0$, which corresponds to a growing size of the extra space. Some fluctuations are created in the range between points A and B, where the kinetic term is negative. Then the field ϕ, being phantom, moves towards a growing potential, to point B. Evidently, point B is an attractor without being a minimum of the potential. A universe for which this situation takes place has a nontrivial basic state, which should be discussed in more detail. To

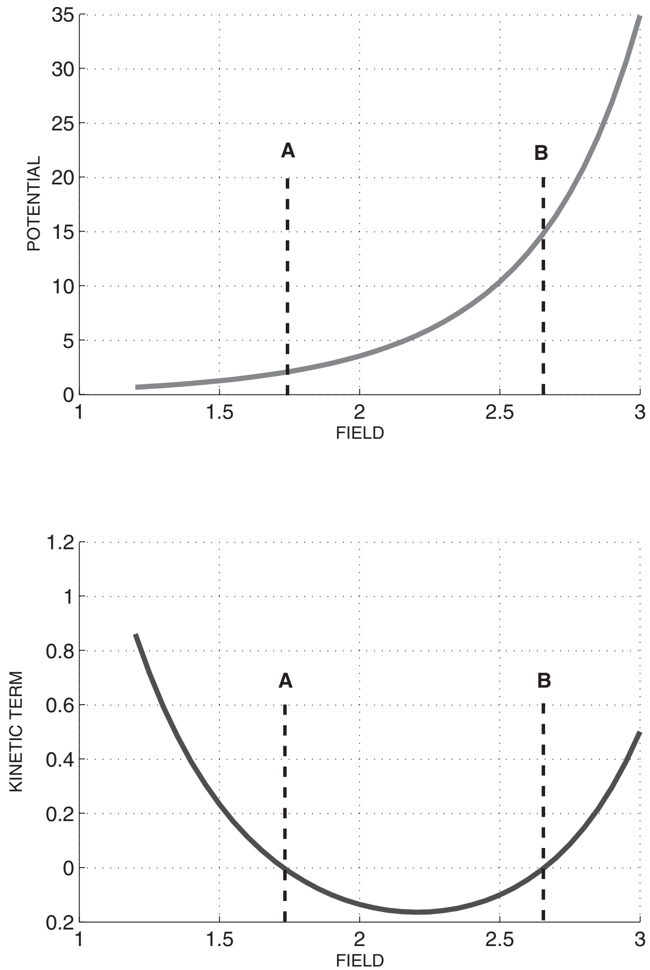


Figure 10.10 The potential (10.132) and the coefficient in the kinetic term (10.133) for the theory (10.131) and $F(R)$ from (10.130). The parameter values: $d = 3$, $c = -0.12$, $\Lambda = 2$, $c_1 + c_2 = -5.5$, $c' = -0.626$.

this end, consider a simplified situation:

$$S = \int d^4x \left[\frac{1}{2}K(\phi)(\partial\phi)^2 - V(\phi)\right], \tag{10.134}$$

with $K(\phi_{\text{crit}}) = 0$ and, without loss of generality, put $\phi_{\text{crit}} = 0$. Then the behavior of $K(\phi)$ and $V(\phi)$ at small ϕ is $K(\phi) = k\phi + o(\phi)$, and

$V(\phi) = V(0) + h\phi + o(\phi)$, where it is assumed that $k > 0, h > 0$, referring to point B in Fig. 10.10. Evidently, near the critical point the field tends to it irrespective of where its initial value is chosen, to the left or to the right of it. Nevertheless, a classical motion near the critical point $\phi = 0$ is absent. Indeed, the classical equation for ϕ in curved space-time has the form

$$K(\phi)\Box\phi + \frac{1}{2}K_\phi(\partial\phi)^2 = -V_\phi, \tag{10.135}$$

or, according to the above expressions for K and V at small ϕ,

$$k\phi\Box\phi = -\left(h + \frac{1}{2}k(\partial\phi)^2\right) + o(\phi). \tag{10.136}$$

Let $(\partial\phi)^2 > 0$ (as in the cosmological case where $\phi = \phi(t)$). Then the right-hand side of Eq. (10.136) is smaller than $-h < 0$ and does not tend to zero as $\phi \to 0$. Hence the second-order derivative in $\Box\phi$ tends to infinity because $\phi \to \phi_{\rm crit} = 0$. Introduction of additional terms with higher-order derivatives into the Lagrangian does not improve the situation. Indeed, in a stationary state, where all derivatives are equal to zero, the classical equation (10.136) with $h > 0$ has no solution. It means that the kinetic energy of a homogeneous basic state cannot be equal to zero.

Multiple factor spaces and a spatially varying size of the extra dimensions

Recall that *a priori* we do not assume a fixed number of extra dimensions, their geometry, and topology. All this emerges at the Planck scale due to quantum fluctuations. The more complex the emerging structure of the extra space, the richer the possibilities.

Consider a space with the metric (10.84) and two extra factor spaces: $\mathbb{M} = \mathbb{M}_4 \times \mathbb{M}_{d_1} \times \mathbb{M}_{d_2}$. Using the same action (10.131), we must now introduce two scalar fields in order to describe the low-energy limit of the theory. The effective potential in the Einstein picture has the form

$$V_{\rm Ein}(\phi_1, \phi_2) = -\frac{1}{2}\,\mathrm{sign}(F'(d_1\phi_1 + d_2\phi_2))\frac{|\phi_1|^{d_1/2}}{[(d_1-1)]^{d_1/2}}\frac{|\phi_2|^{d_2/2}}{[(d_2-1)]^{d_2/2}} \times \frac{F(d_1\phi_1 + d_2\phi_2) + d_1\phi_1^2\left[c_1 + \dfrac{2c_2}{d_1-1}\right] + d_2\phi_2^2\left[c_1 + \dfrac{2c_2}{d_2-1}\right]}{[F'(d_1\phi_1 + d_2\phi_2)]^2}, \tag{10.137}$$

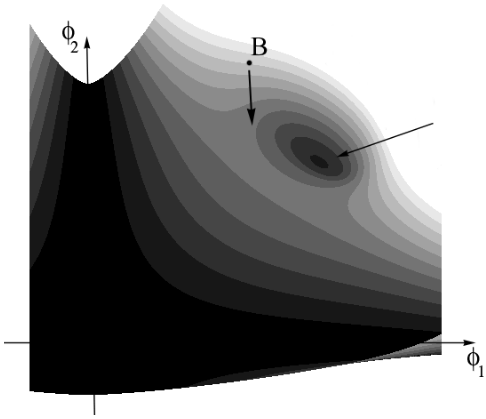


Figure 10.11 The effective potential for a model with the extra space $\mathbb{M}_{d_1} \times \mathbb{M}_{d_2}$, $d_1 = d_2 = 3$, with the parameters $c = -0.5$, $\Lambda = 0.2$, $c_1 = c_2 = -0.38$. View from above, lower levels look darker. A local minimum is marked by a long arrow.

Figure 10.11 presents a potential with two valleys along the mutually perpendicular directions $\phi_1 = 0$ and $\phi_2 = 0$, each of them corresponding to an infinitely large size of one of the factor spaces $\mathbb{M}_{d_1}$ or $\mathbb{M}_{d_2}$. Of greater interest is the local minimum, where both factor spaces have a finite size. The Universe can be located for a sufficiently long time in this metastable state, as in the case with one extra space with a simpler structure discussed above.

An opportunity of interest emerges if we suppose that the Universe was born at point B in Fig. 10.11. During inflation, the field moves from point B along the arrow, and quantum fluctuations create different field values in causally disconnected domains. The fate of a domain depends on the field value inside it. Even if the majority of domains get into the metastable minimum at the end of inflation, some part of them rushes to one of the valleys. In this case our Universe must contain a certain number of domains with a macroscopically large extra space. Their number and sizes strongly depend on the initial conditions.

If in our Universe there are domains with a large extra space, then the laws of low-energy physics in such domains are different from those we know. Thus, if a star flies into such a domain, the balance of forces inside it will be violated, and the star will either collapse or explode.

Thus we have shown that the same basic theory can lead to drastically different low-energy theories, depending on the structure of extra dimensions and the initial data.

10.3.6 Conclusions

It is clear that pure multidimensional gravity, without invoking any fields of nongeometric origin, and even with a very simple choice of the geometry, can be an arena for quite complicated phenomena. It is a result of the curvature-nonlinearity of gravity, but such a nonlinearity inevitably emerges due to quantum corrections to the Einstein theory and must in principle be included in the theory of gravity.

The slow-change approximation, introduced here for multidimensional gravity, substantially simplifies the analysis. This approximation turns out to be valid in a wide range of phenomena at which the curvature and the energy scale are far from Planckian ones (though, Planckian in the multidimensional sense, these scales can appreciably differ from the four-dimensional ones).

On reduction to four dimensions, nonlinear multidimensional gravity leads to a sophisticated scalar field dynamics, with stable and metastable minima of the potential and a nontrivial form of the kinetic term. Some such minima look promising in the sense of describing the modern Universe evolution, with an accelerated expansion of the external space and stable extra dimensions. The latter leads to constant values of the fundamental constants or their extremely slow drift as is the case if the system is only approaching its stable state. In other viable models the extra dimensions are variable, but the variations are small and do not leave the bounds of modern experimental constraints.

Among the unsolved problems are the choice of the physical conformal frame and the necessity of a fine tuning of the parameters for obtaining the observed small value of the cosmological constant.

10.4 The origin of gauge symmetries and fundamental constants

Despite the advances made in fundamental physics, the existence of a large number of theories and approaches suggests that there are problems in describing observational and experimental data. The idea of compact extra dimensions allows one to effectively explain a considerable number of phenomena and indicates the direction of further development in the theory in spite of the absence of direct experimental confirmation of their existence.

Here we will study the problems associated with emergence of gauge symmetries and fundamental constants at the early stage of the evolution of the Universe using the approach proposed in our previous works [83, 358]. This approach is based on the assumption that there are extra dimensions forming a compact space with the properties determining the observed low-energy physics. A solution to the first problem within the Kaluza–Klein approach is well known: gauge symmetries arise as a consequence of the corresponding symmetries of the extra space (see, e.g, [39]). Therefore, the problem is reduced to justification of the choice of symmetric spaces among a set of extra spaces with an arbitrary geometry. More precisely, we assume that, at some instant of time on the Planck scale, the four-dimensional Riemannian space M_4 [196, 402, 427] arises as a result of quantum fluctuations. Simultaneously, there appears a compact extra d-dimensional space $\mathbb{M}_d$. The set of possible geometries of the extra space $\mathbb{M}_d$ is at least a continuum. However, a geometry with a high degree of symmetry is observed because the existence of gauge symmetries is beyond question. The probability of its emergence is negligible, and hence there should exist a selection mechanism that separates the appropriate geometries. One of possible variants based on statistical considerations will be analyzed below.

The second problem discussed in our work is related to the fundamentality of the parameters $\hbar$ and G [147] and consists in the following. The geometric approach to the theory implies the presence of only one scale of the dimension of length ℓ. The parameters of the initial Lagrangian constructed only from the metric tensor are proportional to ℓ^n, where n is an integer. The question arises on the instant of time at which the Planck and gravitational constants appear as independent fundamental parameters and on their relationship with the initial parameters of the theory. An answer to this question for the gravitational constant is well known and, as a rule, is presented in the form

$$M_P^2 = m_D^{D-2} V_d. \tag{10.138}$$

Here, M_P is the Planck mass, m_D is a parameter with the dimension of mass or, more precisely, inverse length, and $V_d < \infty$ is the volume of the extra space $\mathbb{M}_d$. The similar problem of fundamentality of the Planck constant $\hbar$ is less well understood, even though some interesting works have appeared in this direction. In particular, Volovik [411] discussed the place of the Planck constant in modern theory, analyzed different methods of including this constant in the field equations, and considered the consequences of a hypothetic spatial dependence of the Planck constant. However, a specific implementation of these hypothetical variants was not given.

Below, we propose a mechanism of emergence of the Planck constant together with the gravitational constant without a detailed discussion of the possibilities this entails. As a result, we can reveal the relationships of the constants $\hbar$ and G determined from low-energy experiments with the parameters of the initial Lagrangian describing multidimensional gravity with higher derivatives.

Continuing the ideology developed in [358], we assume that spaces with an appropriate geometry arise in space-time foam with some (maybe small) probability. We are interested in spaces with the geometry of a direct product

$$\mathbb{M}_D = \mathbb{M}_{D_1} \times \mathbb{M}_{d_1}, \quad D_1 \geq 4, \;\; d_1 \geq 2 \tag{10.139}$$

and the volumes $V(D_1) \gg V(d_1)$. Transitions with changes in the geometry are conveniently described in terms of the path integral technique [196, 402]. For this purpose, the superspace $\mathcal{M}_D = (\mathbb{M}_D; g_{ij})$ is defined as a set of metrics g_{ij} in the space $\mathbb{M}_D$ up to diffeomorphisms. On a spacelike section Σ we introduce the metric h_{ij} (for details, see [427]) and define the space of all Riemannian $(D-1)$-metrics in the form

$$\text{Riem}\,(\Sigma) = \{h_{ij}(x)\big|x \in \Sigma\}.$$

The transition amplitude from the section Σ_{in} to the section Σ_f is an integral over all geometries allowable by the boundary conditions:

$$A_{f,\text{in}} = \langle\, h_f, \Sigma_f | h_{\text{in}}, \Sigma_{\text{in}} \,\rangle = \int_{h_{\text{in}}}^{h_f} Dg \exp[iS(g)]. \tag{10.140}$$

The absence of the Planck constant in the exponential function is a result of choosing the appropriate units of measurement. Nonetheless, it will be shown below that the Planck constant naturally emerges after the definition of the dimensional units simultaneously with the gravitational constant G.

10.4.1 Why is the extra space symmetric?

The transition amplitude (10.140) generating a four-dimensional space is, as a rule, calculated under the assumption of a special form of the metric h_f on the hypersurface Σ_f with an interval of the type [403, 427]

$$ds^2 = \sigma^2[N(t)^2 dt^2 - a(t)^2 d\Omega_3^2], \quad \sigma^2 = \frac{1}{12\pi^2 M_{\text{Pl}}^2}. \tag{10.141}$$

The transition amplitude (10.140) describes the birth of a space of type (10.139), where $D_1 = 4$ and there is no preferred geometry for the subspace

$\mathbb{M}_{d_1}$. In this section, we propose a symmetrization mechanism for compact spaces. It was shown that the space with a metric that weakly differs from the metric written above asymptotically tends to (10.141) in the course of cosmological expansion.

Let us postulate that the evolution of any closed system is accompanied by an increase in its entropy. It should be noted that the entropy of a subsystem can decrease. In particular, the Hawking evaporation of a black hole decreases its entropy but increases the entropy of the Universe as a whole. Below, we will consider a space with the symmetry (10.139) under the assumption that the subspace volume is considerably larger than the volume of the compact subspace $\mathbb{M}_{d_1}$.

Entropy of a compact space

First and foremost, we will demonstrate that the entropy of a compact space reaches a minimum on a class of maximally symmetric spaces. The initial definition of the entropy in our work coincides with the known definition by Boltzmann, who related the entropy to the number Ω of states of a system: $s = k_B \ln \Omega$. A discussion of other possibilities can be found, for example, in [89, 223]. The notion of the number of states is correctly defined at a quantum level, where the set of energy levels and the degree of their degeneracy are known. However, quantization of the space $\mathbb{M}_d$ means quantization of a gravitational field, which by itself remains an unsolved problem. In this respect, we will restrict ourselves to calculating the number of states from a classical viewpoint. It is sufficient because we are interested in relative quantities (see also [381]).

The entropy of an arbitrary compact space $\mathbb{M}_d$, in which matter fields are absent, is a functional $s[G]$ of the metric tensor G. The number of states is determined by an observer outside the system. It is intuitively clear that the higher the symmetry of an object (in our case, a compact space), the smaller the statistical weight of the object. Let us prove this statement. To calculate the statistical weight Ω, the space $\mathbb{M}_d$ is embedded in the space $\mathbb{R}^N$, which can be done if $\mathbb{M}_d$ is sufficiently smooth and N is sufficiently large [149]. It is assumed that the external observer is located in the space $\mathbb{R}^N$. Each point of the space $\mathbb{M}_d$ is described by the internal coordinate $y \in \mathbb{M}_d$ and the external coordinate $x \in \mathbb{R}^N$. We choose a specific point $P \in \mathbb{M}_d$ and fix its coordinate $x_P \in \mathbb{R}^N$. Then, we choose the set of basis vectors $e_k,\ k = 1, 2, \ldots, N$ in $\mathbb{R}^N$. We require that all d tangent linearly independent vectors at the point P, $e_a^{(P)},\ a = 1, 2, \ldots, d$, which form the coordinate basis in $\mathbb{M}_d$, should be included in this set. Using this basis, we calculate the components of the metric tensor $G_{ab}^{(P)}(x_P)$ of the space $\mathbb{M}_d$. We fix the tangent space $T(P)$ spanned by the vectors $e_a^{(P)}$.

Then we choose another point $Q \in \mathbb{M}_d$, with the coordinates $x_Q \in \mathbb{R}^N$ and the tangent space $\mathbb{T}(Q)$. By moving along the curve $l_{PQ} \subset \mathbb{M}_d$, the point Q is displaced to the point with the coordinates x_P, so that the tangent space $\mathbb{T}(Q)$ coincides with the tangent space $\mathbb{T}(P)$.

The new components of the metric tensor $G_{ab}^{(Q)}(x_P)$ of the space $\mathbb{M}_d$ are calculated at the same point $x_P \in \mathbb{R}^N$. If the metric tensors do not coincide, $G_{ab}^{(Q)}(x_P) \neq G_{ab}^{(P)}(x_P)$, the observer in the space $\mathbb{R}^N$ will fix a new "microstate" and increase the statistical weight of the space $\mathbb{M}_d$ by unity. If the equality

$$G_{ab}^{(Q)}(x_P) = G_{ab}^{(P)}(x_P), \tag{10.142}$$

is valid, the number of microstates remains unchanged. It should be noted that the condition (10.142) corresponds to the existence condition of a Killing vector along the curve PQ. Therefore, the presence of Killing vectors decreases the statistical weight of a compact space. Maximally symmetric spaces have a minimum entropy.

Decay of excitations of a compact space

The compact subspace $\mathbb{M}_{d_1}$ can be considered as a subsystem of the space $\mathbb{M}_D$. Now, we show that, if the subspace volumes satisfy the inequality $V_{D_1} \gg V_{d_1}$, there is an entropy flow from the subspace $\mathbb{M}_{d_1}$ to the subspace $\mathbb{M}_{D_1}$, the entropy of $\mathbb{M}_{d_1}$ tends to a minimum, and its geometry tends to a maximally symmetric one. Furthermore, the entropy of the entire system, i.e., the space $\mathbb{M}_D$ increases.

As a rule, the entropy of a closed system changes along with conservation of the system energy. Since gravitational energy depends on the space topology, we will restrict our consideration to transitions without its changes. In [83, 190], it has been demonstrated that, in the framework of gravity nonlinear in the Ricci scalar, at least local minima of the energy density exist. In this case, there is a set of energy levels, and each geometry of the extra space can be represented in the form of an expansion in eigenfunctions of the d'Alembert operator of the background metric. In terms of the Kaluza-Klein theory, the eigenvalues of this operator contribute to the mass of excitations, which are interpreted as particles in the space $\mathbb{M}_4$. For a compact space with the geometry of a circle of radius r, the mass of the lightest particle is $m_1 = 1/r$, and it is experimentally bounded below by a value of several TeV. It is evident that its decay into light particles propagating in our space should be accompanied by an entropy increase. In the process, the geometry of the extra space relaxes to a state that has a minimum entropy and is characterized by the absence of excitations.

As an illustration, we consider a space of the type (10.139), where $\mathbb{M}_{D_1} = \mathbb{M}_4 \times \mathbb{M}_{d_2}$, $\mathbb{M}_{d_2} = \mathbb{S}_1$, $\mathbb{M}_{d_1} = \mathbb{S}_1$ with the metric g_{MN}, which differs from the diagonal metric $\eta_{MN} = \text{diag}(1, -1, -1, -1, -r_1, -r_2)$ only slightly, so that $g_{MN} = \eta_{MN} + h_{MN}(x, y_1, y_2)$. The size of one of the extra spaces is considerably larger than the size of the other extra space: $r_2 = \beta r_1, \quad \beta \gg 1$. The dynamic equation for the field h_{MN} can be written in the form [39]

$$\eta^{AB}\partial_A\partial_B h_{MN}(x, y_1, y_2) = 0.$$

By substituting the field h in the form of the expansion in eigenfunctions of the d'Alembert operators $Y_n(y_1)$, $Y_n(y_2)$ of both subspaces (in our case, circles),

$$h_{MN}(x, y_1, y_2) = \sum_{n_1, n_2} h_{MN}^{(n_1,n_2)}(x) Y_{n_1}(y_1), Y_{n_2}(y_2),$$

we obtain the equation for the components

$$\left(\Box_x + \frac{n_1^2}{r_1^2} + \frac{n_2^2}{r_2^2}\right) h_{MN}^{(n_1,n_2)}(x) = 0. \tag{10.143}$$

The microstates of the subspaces $\mathbb{M}_{d_1}$ and $\mathbb{M}_{d_2}$ are characterized by the integers n_1 and n_2. The first excited state of the subspace $\mathbb{M}_{d_1}$ has a mass $m_1 = 1/r_1$. A decay of this state into two excited states of the subspace $\mathbb{M}_{d_2}$ with masses $m_2 = n_2/r_2$ and $m_2' = n_2'/r_2$, $n_2, n_2' = 0, 1, 2, \ldots, [\beta]$ and zero momenta occurs with energy conservation,

$$m_1 = m_2 + m_2' \tag{10.144}$$

and an increase in the entropy of the entire system. This increase is associated with the fact that the number of different microstates of the space $\mathbb{M}_{d_2}$ that satisfy the condition (10.144) is of the order of $\Omega_2 = \beta/2 \sim r_2/r_1$. In this case, the geometry of the subspace $\mathbb{M}_{d_1}$ tends to a maximally symmetric geometry because the number of excitations in the subspace tends to zero and the entropy tends to a minimum.

It is obvious that the number of microstates increases with an increase in the volume of the subspace $\mathbb{M}_{d_2}$. Inclusion of states with different four-momenta in the Minkowski space $\mathbb{M}_4$ substantially enhances the effect.

Let us increase the dimension of the compact space with a large volume and consider a manifold of the type (10.139) where

$\mathbb{M}_{D_1} = \mathbb{M}_4 \times \mathbb{M}_{d_2}$, $\mathbb{M}_{d_2} = \mathbb{S}_2$, $\mathbb{M}_{d_1} = \mathbb{S}_1$. It is known that the excitation mass in $\mathbb{S}_2$ is $m(l) = \sqrt{l(l+1)}/r_2$, and the degeneracy factor is $2l+1$. The decay of an excitation of the space $\mathbb{S}_1$ into two excited states of the space $\mathbb{S}_2$ also proceeds with energy conservation: $m_1 = m(l) + m(l')$. It is easy to see that the number of allowed energy states with zero momentum is also proportional to the volume of the large space: $\Omega_2 \simeq \beta^2 = r_2^2/r_1^2$.

Therefore, if quantum fluctuations generate a space of the type $\mathbb{M}_a \times \mathbb{M}_b$, the entropy flow is directed toward the subspace of larger size. The geometry of the subspace of smaller size tends to a maximally symmetric geometry compatible with its topology. The existence of gauge symmetries in the main space appears to be a purely statistical effect.

The transition rate to a symmetric state

The entropy of the entire system consisting of two subspaces (10.139) increases as a result of transformations of particles (excitations) of the compact subspace $\mathbb{M}_{d_1}$ to particles propagating in the subspace M_{D_1}. The compact extra space that, at the initial time instant, had an arbitrary geometry, acquires a maximally symmetric shape during the entropy transfer to the main space. In this case, the entropy of the entire system (including the extra and main spaces) increases while the entropy of a subsystem (the compact extra space) tends to a minimum. The situation resembles the third law of thermodynamics, according to which the entropy of a body tends to zero with a decrease in the thermostat temperature.

The subspace $\mathbb{M}_{D_1}$ in (10.139) does not necessarily have a direct-product structure. The main requirement to this subspace is the presence of a large number of energy levels that contribute to the statistical weight of the system at fixed temperature. The Minkowski space M_4 itself exhibits this property to a full extent.

Let us estimate the "symmetrization" rate of the extra space. Weak deviations of the geometry from an equilibrium configuration can be interpreted as excited states of mass m_1 (see, e.g.,[12]). Since it is the only scale, it should be expected that the probability of decay will satisfy the relation $\Gamma \sim m_1 \sim 1/L_d$, where L_d is the characteristic size of the extra space. Setting $L_d \leq 10^{-17}$ cm, we find that the lifetime of the excited state is $t_1 \sim L_d \leq 10^{-27}$ s. Therefore, the extra space transformed into the most symmetric state long before the onset of the primordial nucleosynthesis but maybe after completion of the inflationary stage. However, states that correspond to the first excited level of the Kaluza-Klein tower with a lifetime of the order of 10^5 s or more are also considered [345]. Under these conditions, the theory acquires gauge invariance well after the nucleosynthesis stage, which creates new possibilities and problems. In order that a Kaluza-Klein particle be stable, it is necessary to make additional assumptions that complicate the structure of the extra space [341].

10.4.2 Fundamental constants and the properties of an extra space

Now we will discuss the problem associated with the emergence of physical parameters in the modern Universe. A more detailed discussion is presented in Sec. 11.2. In a consistent implementation of the Kaluza-Klein idea, boson fields represent individual metric components of the extra space. The only scale possible in this situation is the length scale ℓ; however, it is hard to use even this unit because quantum fluctuations in the space-time foam continually renormalize any parameters in an unpredictable manner. The situation becomes better as soon as a classical spacetime region arises due to the same quantum fluctuations. It is also accompanied by the appearance of independent stationary quantities that can be chosen as measurement units, for example, the volume of the compact space V_{d_1} and the vacuum energy density U_m. Only in this case can dimensional physical constants, including fundamental constants, be fixed. A possible variant of this scenario will be considered below.

The action is chosen in the form (see, e.g., [310, 377, 386]),

$$S = N_0 \int d^D y \sqrt{-G} F(R; a_n). \tag{10.145}$$

The parameters a_n, N_0 take on specific values in the birth of this spatial region [358] after the choice of the length unit. The appropriate choice of the parameters a_n can provide boundedness of the effective action from below [77, 83]. The dimensionality of the arbitrary function $F(R)$ is conveniently chosen to coincide with the dimension of R, i.e., it is l^{-2}. Then, the dimension of N_0 is l^{2-D}. Note that the normalization constant N_0 equals $1/(16\pi G)$ only in the low-energy limit.

The metric of the space $\mathbb{M}_D$ is written in the form [83, 101]

$$\begin{aligned} ds^2 &= G_{AB} dX^A dX^B = g_{ab}(x) dx^a dx^b - \mathrm{e}^{2\beta(x)} \gamma_{ij}(y) dy^i dy^j \\ &= \mathrm{N} dt^2 - g_{\mu\nu}(x) dx^\mu dx^\nu - \mathrm{e}^{2\beta(x)} \gamma_{ij}(y) dy^i dy^j. \end{aligned} \tag{10.146}$$

Here g_{ab} is the metric of the subspace $\mathbb{M}_4 = \mathbb{R} \times \mathbb{M}_3$ with the signature $(+---)$, $e^{2\beta(x)}$ is the curvature radius of the compact subspace $\mathbb{M}_d$, and $\gamma_{ij}(y)$ is its positive-definite metric. For a specified foliation of space by spacelike surfaces, we can always choose the normal Gaussian coordinates which are used in the last equality in the expression (10.146). We assume that the coordinates x, y have the dimension l, time is measured in seconds, and the lapse function N is an unknown parameter with the dimension of $(\ell/s)^2$. The metrics g_{ab} and γ_{ij} are dimensionless.

According to [77, 83], the topology and metric on the spacelike section Σ_f of amplitude (10.140) are determined by imposing the following conditions:

(i) The topology of $\mathbb{M}_D$ has the form of a direct product,

$$\mathbb{M}_D = \mathbb{M}_4 \times \mathbb{M}_d, \tag{10.147}$$

where d is the dimension of a compact extra space.

(ii) The curvature of the subspace $\mathbb{M}_d$ satisfies the condition

$$R_4(g_{ab}) \ll R_d(\gamma_{ij}). \tag{10.148}$$

(iii) As shown above, the extra space with an arbitrary geometry evolves to a space with a maximum number of Killing vectors possible for the topology under consideration. In this respect, in the set of subspaces $\mathbb{M}_d$ we choose maximally symmetric spaces of constant curvature R_d, related to the curvature parameter k in a typical manner,

$$R_d(\gamma_{ij}) = e^{-2\beta(x)} kd(d-1)\ell^{-2}. \tag{10.149}$$

Owing to the special form of the chosen metric (10.146), the following relations hold, see Sec. 10.2.2:

$$\begin{aligned} R &= R_4 + \phi + f_{\text{der}}, \\ \phi &= kd(d-1)\, e^{-2\beta(x)} l^{-2}, \\ f_{\text{der}} &= 2dg^{\mu\nu}\nabla_\mu\nabla_\nu\beta + d(d+1)g^{\mu\nu}\partial_\mu\beta\partial_\nu\beta, \end{aligned} \tag{10.150}$$

where the field ϕ is defined by explicitly introducing its dimension l^{-2}. The covariant derivative ∇ acts in the space $\mathbb{M}_4$. The volume $\mathcal{V}_d$ of the internal space of unit curvature depends on its geometry, being expressed in terms of its own metric:

$$\mathcal{V}_d = \int d^d y \sqrt{\gamma}. \tag{10.151}$$

In what follows, we will use the slow-change approximation (see Sec. 10.2.2)

$$|\phi| \gg |R_4|, \quad |f_{\text{der}}|, \tag{10.152}$$

which is valid even at the onset of the inflationary stage. Then, by expanding the expression $F(R;a_n) = F(\phi + R_4 + f_{\rm der}; a_n)$ in a Taylor series and integrating over the extra-space coordinates we obtain the expression

$$S \simeq \mathcal{V}_d\, N_0 \int \sqrt{{}^4g}\, d^4x\, e^{d\beta} + [F'(\phi;a_n)R_4 + F(\phi;a_n) + F'(\phi;a_n) f_{der}], \qquad (10.153)$$

which is typical of a scalar-tensor theory of gravity in the Jordan frame. Using the conformal transformation

$$g_{\mu\nu} \mapsto \widetilde{g}_{\mu\nu} = |f(\phi)| g_{\mu\nu}, \quad f(\phi) = e^{d\beta} F'(\phi;a_n), \qquad (10.154)$$

we pass on to the Einstein frame:

$$S = \mathcal{V}_d\, N_0 \int d^4x\, \sqrt{\widetilde{g}}\, (\mathrm{sign}\, F')L, \qquad (10.155)$$

$$L = R_{\mathbf{I}} + \frac{1}{2} K(\phi)(\partial\phi)^2 - U(\phi), \qquad (10.156)$$

$$K(\phi) = \frac{1}{2\phi^2}\left[6\phi^2\left(\frac{F(\phi;a_n)''}{F(\phi;a_n)'}\right)^2 - 2d\phi\frac{F(\phi;a_n)''}{F(\phi;a_n)'} + \frac{1}{2}d(d+2)\right], \qquad (10.157)$$

$$U(\phi) = -(\mathrm{sign}\, F(\phi;a_n)')\left[\frac{|\phi|\cdot \mathbf{I}^2}{d(d-1)}\right]^{d/2} \frac{F(\phi;a_n)}{F'(\phi;a_n)^2}, \qquad (10.158)$$

where $F(\phi;a_n)' = dF/d\phi$.

Assuming the existence of a minimum of the potential (10.158) at $\phi = \phi_m$ and using the slow-change approximation, the action (10.155) can be presented near this minimum in the form

$$S \simeq N_0 \mathcal{V}_d c_I \int dt d^3x_I \left[R_I + \frac{1}{2}K(\phi_m)(\partial_I\phi)^2 - U(\phi_m) - \frac{1}{2}U''(\phi_m)(\phi-\phi_m)^2\right], \qquad (10.159)$$

where

$$(\partial_I\phi)^2 = (\partial\phi/c_I\partial t)^2 - (\partial\phi/\partial x_I)^2.$$

Here, we have taken into account that $\sqrt{g} = c_I$ in four-dimensional Minkowski space. The index l designates the used unit of length, and c_I is the speed of light in the chosen units.

The measurement units are related by the expression

$$l = \alpha \cdot \text{cm}. \tag{10.160}$$

where α is an as yet unknown parameter. In Eq. (10.159), we pass over to the standard units of length:

$$S = N_0 \mathcal{V}_d c\alpha^2 \int dt d^3x \left[R_4 + \frac{1}{2} K(\phi_m)(\partial\phi)^2 - \alpha^2 U(\phi_m) \right.$$
$$\left. - \alpha^2 \frac{1}{2} U''(\phi_m)(\phi - \phi_m)^2 \right]. \tag{10.161}$$

Here, $x_I = \alpha x$, $c_I = \alpha c$, $R_I = R_4/\alpha^2$, and $(\partial_I \phi)^2 = (\partial\phi)^2/\alpha^2$.

Since the expression (10.145) is a low-energy limit of the action (10.161), it should adequately describe purely gravitational phenomena. Moreover, the expression (10.161) makes it possible to explain the origin of the inflaton potential and the cosmological constant. The effective action (10.161) contains the initial parameters of the theory without using fundamental parameters, such as the Planck constant $\hbar$ and the gravitational constant G. However, from practical and historical viewpoints, it is more convenient to explicitly introduce these parameters. In terms of the developed approach, this can be done by imposing the constraints

$$N_0 \mathcal{V}_d c\alpha^2 = \frac{c^4}{16\pi G\hbar}, \tag{10.162}$$

$$\alpha^2 U(\phi_m) = \frac{16\pi G}{c^4} \Lambda. \tag{10.163}$$

Here Λ is the observable vacuum energy (dark energy) density. It is important that such quantities as the volume of the extra space $\mathcal{V}_d$ and the minimum value of the potential $U(\phi_m)$ acquire specific values only at $\phi = \text{const}$. If the field ϕ is identified with the inflaton, a nontrivial situation arises. The inflationary stage is completed before the scalar field (inflaton) reaches a potential minimum. Consequently, at the inflationary stage, when the Universe formed and the field ϕ varied with time, the quantities G and $\hbar$ should also have been time-dependent. In this case, the theory of gravity is an effective theory valid at low energies (see also [410]).

Let us introduce the definition

$$m_\Phi \equiv \alpha \sqrt{\frac{U''(\phi_m)}{K(\phi_m)}}, \tag{10.164}$$

and, under the assumption $K(\phi_m) > 0$, change the variables:

$$\sqrt{\frac{c^4}{16\pi G}K(\phi_m)}(\phi - \phi_m) = \Phi.$$

As a result, we arrive at the conventional form of the action for a scalar field Φ interacting with gravity, that is,

$$S = \frac{1}{\hbar}\int dt\, d^3x \left(\frac{c^4}{16\pi G}R_4 + \frac{1}{2}(\partial\Phi)^2 - \Lambda - \frac{1}{2}m_\Phi^2\Phi^2\right). \qquad (10.165)$$

Now, it becomes clear that the parameter m_Φ has the meaning of the scalar field mass.

Equations (10.162) and (10.163) allow us to solve the problem of separate emergence of the gravitational and Planck constants, which appear to be functions of the parameters of the theory N_0, $\{a_n\}$. The set of parameters $\{a_n\}$ is also implicitly included in the expressions for the quantities $\mathcal{V}_d, U(\phi_m)$. We do not consider here the origin of the lapse function and hence the speed of light. Eliminating the insignificant parameter α from Eqs. (10.162), (10.163) and (10.164), we obtain

$$\frac{U(\phi_m)K(\phi_m)}{U''(\phi_m)} = \frac{16\pi G}{c^4}\frac{\Lambda}{m_\Phi^2}, \qquad (10.166)$$

$$N_0\mathcal{V}_d(\phi_m)\frac{K(\phi_m)}{U''(\phi_m)} = \frac{c^3}{16\pi G\hbar m_\Phi^2}. \qquad (10.167)$$

Here, the right-hand sides contain measurable quantities whereas the left-hand sides involve quantities depending on the initial parameters of the theory. At energies on the Planck scale, when the field ϕ did not reach the potential minimum, the use of the modern values of the fundamental constants G and $\hbar$ requires some care. Moreover, assuming that the vacuum energies Λ are different in different regions of the Universe [142, 170, 358], the fundamental constants G and $\hbar$ are also different according to the expressions (10.166), (10.167). A discussion of this problem and references can be found in [411].

At the inflationary stage, when ϕ varied, the quantities G and $\hbar$ also varied. The dependence of these quantities on the field can be determined from (10.166) and (10.167) through the change $\phi_m \to \phi$, which are true for small deviations of the field from the equilibrium position.

As a result, we have

$$G = G(\phi) \simeq \frac{c^4 m_\Phi^2}{16\pi\Lambda} \frac{U(\phi)K(\phi)}{U''(\phi)},$$
$$\hbar = \hbar(\phi) \simeq \frac{c^3}{16\pi G(\phi)\mathcal{V}_d(\phi) N_0}. \tag{10.168}$$

The time dependence of the inflaton field ϕ in different inflationary models is well understood [282], hence the dynamics of the fundamental parameters can be described by choosing a specific inflationary model.

The conversion factor between the units of length $\alpha = l/cm$ can also be found by another method. Indeed, the characteristic size of the extra space is determined to be $\mathcal{V}_d^{1/d}$ in units of l. Moreover, by designating the size of the hypothetical extra space as L_d, expressed in terms of cm, we obtain $\alpha = L_d/\mathcal{V}_d^{1/d}$. The values of $L_d \leq 10^{-17}$ do not contradict the experimental data. With due regard for Eq. (10.164), we find the constraint on the initial parameters

$$m_\Phi \mathcal{V}_d^{1/d} \sqrt{\frac{K(\phi_m)}{U''(\phi_m)}} = L_d < 10^{-17}. \tag{10.169}$$

Equations (10.166), (10.167) and (10.169) allow us to determine the parameters of the theory $N_0 a_n$ (see Eq. (10.145)) from the known constants $c, \hbar, G$ and the energy density Λ, the mass of the scalar field m_Φ, and the volume of the extra space $\mathcal{V}_d$. Implementation of this program is a matter for the future. Actually, the fundamental constants are measured with high accuracy, whereas the vacuum energy density is determined with a considerably lower accuracy. It is especially true regarding the mass of the hypothetic scalar field, even though this field is identified with the inflaton. The greatest uncertainty is associated with the volume of the extra space, which is yet to be revealed.

Chapter 11

The emergence of physical laws

11.1 Fine tuning of the Universe parameters

One of the impressing observational facts is the fine tuning of the Universe parameters. This means that all microscopic parameters, the particle masses and coupling constants, not only have the values at which an intelligent life can appear in the Universe, but the admissible values of these parameters are confined to very narrow ranges. Tiny deflections from their real values would lead to catastrophic consequences for the existence of complicated structures. Before proceeding with the discussion, let us consider a few examples.

The electron and the properties of the Universe

The characteristic masses of fermions are of the order of 1 GeV. Only the electron has a mass three orders of magnitude smaller. Imagine that its mass became a few times larger. It could seem that nothing significant should happen, all the same it is still a very light particle. However, an electron mass increase by a factor of three would lead to catastrophic consequences for the Universe. Indeed, the mass difference between the neutron and the proton is $\Delta m = m_n - m_p = 1.28$ MeV, whereas the mass of a "new electron" amounts to $m_E = 3m_e = 1.53$ MeV. It means that the neutron becomes stable because its decay via weak interaction

$$n \to p + e + \bar{\nu}_e$$

is energetically forbidden. Moreover, the allowed reaction

$$p + e \to n + \nu_e,$$

will destroy the stability of hydrogen atoms. Consequently, immediately after the recombination period, practically all protons and electrons are converted to neutral particles. Charged particles simply disappear long before the formation of the first stars. In this case stars will not be formed at all since the gravitational contraction of clouds is efficient only if there is an energy dissipation process.

Thus, an insignificant growth of the electron mass leads to the absence of stars in the Universe and consequently the absence of any, to say nothing of intelligent, life.

The carbon level

Fine tuning works so unconditionally that with its aid one can even estimate unknown parameters. The most well-known example is Fred Hoyle's prediction.

Carbon in stars is formed in two stages. At first two alpha particles merge, forming the unstable isotope beryllium-8. Then a third alpha particle joins this beryllium, forming a carbon nucleus. But beryllium-8 rapidly decays. Therefore, in forming the Universe, it would be reasonable to provide a resonant interaction between beryllium-8 and an alpha particle. Nature did precisely that: the energy level equal to 7.65 MeV is remarkable in that the energy of an excited state of the carbon nucleus is only 0.3 MeV higher than than the summed mass of the alpha particle and the beryllium nucleus. This 0.3 MeV is compensated by the kinetic energy of colliding particles, resonantly increasing the efficiency of the reaction. It is these considerations that led Fred Hoyle to his theoretical prediction of the energy of this level in 1953. When our Universe was only forming, Nature must have already "known" about the future necessity of this level for the existence of carbon life.

Slow reactions in stars

Let us mention one more interesting example of fine tuning of the Universe parameters. Consider the question: what should the Fermi constant G_F be for the emergence of our type of intelligent life? Experience tells us that it requires about five billion years for that. It is the time — not less than that — for which nuclear reactions should last in a star like the Sun. Thus the time of the first reaction in the proton-proton cycle,

$$p + p \to^2 H + e^+ + \nu_e \tag{11.1}$$

should be $t = 1/(\sigma v n) \sim 5 - 10$ billion years, where $v \sim \sqrt{2T/m_p}$ is the mean proton velocity in a star, while the mean concentration of protons inside a star is $n \sim 1.5\,\mathrm{g/cm}^3$. Nuclear reactions begin at temperatures $T \approx 10^7$. The cross-section of the first reaction of the proton-proton cycle, occurring due to the weak interaction, is $\sigma \sim G_F^2 E^2 \sim G_F^2 T^2$.

Collecting all that, we obtain an expression for the Fermi constant

$$G_F \sim \frac{m^{1/4}}{(nt)^{1/2} T^{5/4}}, \tag{11.2}$$

connecting it with such parameters as the mean stellar concentration of protons and temperature, the proton mass and the intelligent life emergence time, or, more precisely, the stellar lifetime. Substituting the numerical values, we obtain $G_F \sim 10^{-5}\,\mathrm{GeV}^{-2}$, which coincides by order of magnitude with the experimental value.

Let us note that nuclear reactions usually occur rapidly. The small value of the Fermi constant makes it possible to slow down the stellar nuclear processes, beginning with the reaction (11.1). Let us look how efficient this slowing-down is. The solar luminosity is known: the energy per second released by the Sun is $L_\odot = 3 \cdot 10^{26}$ J/s. Consequently, 1 kg of solar matter emits

$$\varepsilon_\odot \sim L_\odot / M_\odot \sim 10^{-7}\,\mathrm{J/(s \cdot kg)}.$$

Compare this with the specific energy release of a human being of 80 kg:

$$\varepsilon_{\text{homo}} \sim \frac{2000 \cdot 10^3 \cdot 0.01 \cdot 4.18}{24 \cdot 3600 \cdot 80} \sim 0.02\,\mathrm{J/(s \cdot kg)}.$$

Here, for the estimation, it is assumed that only 1% of the energy absorbed by man, 2000 kcal, passes over to heat. Owing to the smallness of the Fermi constant, stellar nuclear reactions turn out to be by many orders of magnitude less efficient as compared with chemical reactions.

Question. *If chemical reactions inside a human being occur much more rapidly than nuclear reactions in stars, why is the temperature of a human body so much lower than the solar temperature?*

The weak interaction should be really very weak in order that the neutron live long enough. In addition, during Supernova explosions, neutrinos, owing to the weak interaction, are able to carry away the main part of energy from the central part of the star, allowing its inner layers to rapidly contract. But on the other hand, the weak interaction constant cannot be too small. Otherwise, first, the number of neutrinos created in an exploding star would be too small, and second, the outer layers of such a star would not receive sufficient energy from the neutrinos to fly away in space.

Stellar lifetime

Heavy elements are formed in the stars but they must be eventually ejected to space, in order that future stellar generations with planetary systems could form. Therefore the parameters of the theory should be selected in such a way that at least a certain part of the first stellar generation have a short lifetime ending with an explosion. But the next stellar generations should live for a long time (ten billion years) to make possible the emergence of intelligent life. Nature has been able to satisfy these contradictory requirements.

The long life of the next stellar generation rests on the slow proton-proton cycle (here is where the weak interaction is necessary). The short lifetime of the first stars takes place due to large masses plus primordial helium burning. But how to create the primordial helium?

Passage through a needle's eye

To create helium, neutrons are necessary. The latter, created when the Universe was hot, should eventually decay. How to preserve them? Can we preserve neutrons inside deuterium?

$$p + n \to D + \gamma.$$

But the inverse reaction prevents:

$$D + \gamma \to p + n,$$

it is efficient at high temperatures. The binding energy of deuterium is not large, 2.234 MeV, and at temperatures higher than 10^9 K deuterium is rapidly destroyed. The threshold temperature 10^9 K of the Universe is achieved in 300 s after the Beginning. If the neutrons had the "usual" lifetime like 10^{-6} s, their concentration by that time would be vanishingly small.

A way out is a long-lived neutron (900 s)! After this time, the temperature decreases enough, so that the energy of photons becomes insufficient to destroy the deuterium nuclei. The remaining 24 per cent of neutrons turn out to be sufficient.

It is good time to ask a question: what should an Ultimate Theory look like?

It seems that there are two possible answers to this question:

A. The laws of Nature are strictly derived from an initial postulate and simply cannot be different. The whole history of science can be viewed as an approach to this postulate.

B. There are many universes with different laws, and we live in one of them. This approach is gaining more and more adherents, but is it constructive? Suppose that somebody wishes to create a universe like ours, at least purely theoretically. The above examples are designed to convince the reader that it is not a simple task. Indeed, the number of problems to be encountered by the brave creator is immense. Starting from the beginning she (he) must choose properties of all particles and their interactions keeping in mind the future evolution. The sketch looks as follows.

Choosing particles and interactions

The basic components of complex structures are protons and electrons, which already give rise to the simplest of the atoms, the hydrogen atom. More massive nuclei are also necessary, for instance, carbon, the base of organic compounds. But massive nuclei consisting of only protons will immediately decay because of the protons' electromagnetic repulsion. One will have to introduce a new interaction providing attraction. We call it the strong interaction. But then too large nuclei will form, which is also bad, as the matter density formed will be too large. Therefore we make the strong interaction short-range.

The situation has become much better, but the nuclei, though slowly, still decay. If, due to fluctuations, one of the protons flies a little farther from the others, the strong interaction cannot withstand the electric repulsion, the proton escapes, and the nucleus decays. It is evident that the goal is close, only a little is needed to make nuclei heavy enough and stable. The last stroke: we introduce the neutron, having no charge but taking part in the strong interaction. Now the nuclei consist of protons and neutrons and are stable up to the nucleus of iron. All that should have been provided by Nature from the very beginning.

To create complex structures, one needs a large volume, and the expansion solves this problem. But the expansion rate cannot be arbitrary, for too many effects depend on it. The most basic is apparently star formation.

The expansion of space leads to such cooling of the medium at which no intelligence could live. The contradiction can be avoided by creating a locally warmed medium. This function is performed by the stars. Besides, the stars are factories of heavy elements. The long-term existence of the stars is one more problem that the nature had to solve.

It seems evident that the idea of a "multiverse" is a good basis for solving the fine-tuning problem, be it "science" or not.

More examples of fine tuning of the Universe — briefly

We further enumerate the parameters that need to be in a narrow range to make complex structure creation in the Universe possible.

1. The strong interaction constant.
Larger: no hydrogen and light nuclei, e.g., carbon.
Smaller: nuclei heavier than hydrogen are unstable.

2. The weak interaction constant.
Larger: (a) too much primordial helium in the first stars $\Rightarrow$ too large amounts of heavy elements are created in the stars; (b) supernova explosions are absent. The nuclei remain in stellar remnants.
Smaller: (a) too little primordial helium in the first stars; (b) supernova explosions are absent. The nuclei remain in stellar remnants.

3. The gravitational constant.
Larger: the stars are too hot and therefore burn too rapidly (while five billion years are needed to create intelligence).
Smaller: cold stars. Consequently, no nuclear reactions and their product, heavy elements.

4. The electromagnetic interaction constant.
Larger: the atoms are smaller (all structures become more dense). The nuclei are unstable.
Smaller: too weak chemical bonds.

5. The ratio of the electromagnetic interaction constant to the gravitational constant.
Larger: stellar masses are high and therefore their lifetimes are small.
Smaller: no massive stars and therefore no heavy elements.

6. m_e/m_p.
Larger: weak chemical bonds.
Smaller: weak chemical bonds.

7. The Universe expansion rate.
   Larger: galaxies have no time to form.
   Smaller: the Universe collapses before the stars can form.
8. The mean baryonic energy density.
   Larger: too much primordial helium, the stars burn too rapidly.
   Smaller: too little primordial helium, hence too little amounts of heavy elements.
9. Primordial density fluctuations.
   Larger: the Universe would consist of black holes.
   Smaller: the stellar formation process is suppressed.
10. Mean interstellar distance.
    Larger: the heavy element density is insufficient for making solid planets.
    Smaller: unstable planetary orbits.
11. Beryllium-8 decay rate.
    Larger: absence of elements heavier than beryllium.
    Smaller: the nuclear reaction rate increases, leading to small stellar lifetimes.
12. The initial excess of nucleons vs. antinucleons.
    Larger: too large baryon density. Mostly black holes form.
    Smaller: baryonic matter insufficient for stellar formation.
13. Frequency of Supernova explosions.
    Larger: life on the planets is destroyed.
    Smaller: heavy elements insufficient for making solid planets.
14. Dark matter density.
    Larger: an early collapse of the Universe.
    Smaller: galaxy formation suppressed.
15. The neutron–proton mass difference.
    Larger: smaller neutron lifetime, hence too little primordial helium.
    Smaller: rapid neutron decay. Heavy elements are absent.
16. The number of spatial dimensions(3)
    Larger: planets are not kept near stars, no stationary orbits.
    Smaller: hard to provide the viability of complex structures.

11.2 Fine tuning mechanisms

All the above considerations were aimed at convincing the reader that choosing the parameters for forming complex structures in the Universe is a task of formidable difficulty. Admissible ranges of their variation are extremely narrow. Whatever the future theory, the latter statement will refer to it and its parameter set. All that means that the probability of a random implementation of such a parameter set is close to zero. From this viewpoint, the idea of multiple universes with different properties is more attractive. The question is "only" the mechanism of their formation. Multidimensional gravity gives such an opportunity, and that is what is discussed in what follows.

11.2.1 Cascade birth of universes in multidimensional spaces

In this section, the formation mechanism of universes with distinctly different properties is considered in the framework of pure gravity in a space of $D > 4$ dimensions. We also discuss the emergence of the Planck scale and its relationship with the inflaton mass.

The dynamics of our Universe is well described by a modern theory containing 30 to 40 parameters. This number, whose value is determined experimentally, is too large for a theory to be considered as an ultimate one. In addition, it is well known that the range of admissible parameter values is extremely narrow (fine tuning of parameters) for the birth and existence of such complex structures as our Universe, which is hard to explain. Extensive literature is devoted to a discussion of this problem, see e.g. [7]. One way of solving it is based on the assumption of multiple universes with different properties [220, 269, 353]. Rich opportunities of justifying this assumption are contained in the idea of multidimensionality of our space itself. The number of extra dimensions has long been a subject of debate. Thus, the Kaluza-Klein model originally contained one extra dimension. At present, for example, infinite-dimensional spaces [106] and even variable-dimensional spaces [40] are being discussed. The concept of superspace is extended to a set of superspaces with different, unbounded from above numbers of dimensions [358]. Based on the introduced extended superspace, we suggest a formation mechanism of universes with distinctly different properties and the emergence mechanism for the Planck scale. We discuss the probability of quantum transitions that produce lower-dimensional subspaces.

Let us define according to [358] the superspace $\mathcal{M}_D = (\mathbb{M}_D, g_{ij})$ as a set of metrics g_{ij} in the space $\mathbb{M}_D$ up to diffeomorphisms. On a spacelike section Σ, let us introduce a metric h_{ij} (see, e.g., [427] for details) and define the space of all Riemannian $(D-1)$-metrics:

$$\mathrm{Riem}(\Sigma) = \{h_{ij}(x) \,|\, x \in \Sigma\}. \tag{11.3}$$

The transition amplitude from one arbitrarily chosen section Σ_{in} with the corresponding metric h_{in} to another section Σ_f with a metric h_f is written above, (10.140).

The topologies of the sections Σ_{in} and Σ_f can be different. We will be concerned with quantum transitions in which the topology of the hypersurface Σ_f is a direct product of subspaces, $\mathbb{M}_{D-1-d} \times \mathbb{M}_d$. The space $\mathbb{M}_d$ is assumed to be compact.

The entire analysis is performed in the framework of nonlinear gravity in a space of $D > 4$ dimensions without including any matter fields. We discuss the emergence of the Planck scale and its relationship with the inflaton mass. The reduction to a lower-dimensional space is made in several steps to produce a cascade. Different cascades give rise to four-dimensional spaces with different effective theories and different numbers of extra dimensions.

The parameters of the low-energy theory turn out to depend on the topology of the extra spaces and vary in a wide range (see also [118, 165]), though the parameters of the original theory are fixed. This also applies to such fundamental concepts as, for example, the Planck mass and the topology of the extra space. The absence of matter fields postulated here at the outset is a fundamental point. It is suggested that the metric tensor components of the extra (super)space at low energies will be interpreted as matter fields in the spirit of Kaluza–Klein theories.

11.2.2 Simultaneous formation of space-time and the parameters of the theory

In this section we investigate the problems of emergence of the fundamental constants as well as other parameters at an early stage of the Universe's evolution on the basis of the idea of multiple universes. This idea generally implies the existence of an initial Lagrangian with specific parameters and a potential density with numerous local minima. Each of these minima corresponds to a certain low-energy effective theory with its own unique set of parameters. Which of the minima will be ours depends on the initial conditions. Therefore, the observed low-energy physics depends not only on the initial parameters of the Lagrangian but also on the initial metric tensor of the created space-time.

This idea is usually developed in the framework of string theory [391], which incorporates, among other assumptions, the existence of extra dimensions. Here we develop a purely geometric approach [83], postulating only the existence of extra dimensions. We use the cascade reduction mechanism, introduced in [358], to explain the fine tuning of the parameters in our Universe without invoking any string theory assumptions other than the existence of extra dimensions.

The essence of our idea is as follows. Consider a certain multidimensional space. Due to quantum fluctuations, in some of its regions, the geometry of a direct product of two subspaces can arise. Suppose the curvature of one of these subspaces significantly exceeds that of the other; let us refer to the former as the ***extra*** subspace and to the latter as the ***main*** one. Quantum fluctuations in some region of a newly formed main subspace similarly divide it into a direct product of two subspaces. Considering further divisions, we arrive at a "chain" of space partitions. We will call every such partition a ***reduction*** of space. Multiple consecutive reductions will be called a ***reduction cascade***.

Thus a ***reduction cascade*** consists of several steps reducing the effective dimensionality of space. It will be shown that every step of a cascade changes the parameters of the Lagrangian. Therefore, by choosing different cascades we can obtain different final Lagrangians starting from fixed initial parameters. Each Lagrangian corresponds to universes with distinct properties. It is this possibility that is usually associated with the concepts of a landscape, i.e., numerous low-energy vacua. Our goal is, in a sense, the opposite: we try to assess the set of all parameters of the initial Lagrangian leading to the observed fundamental constants.

We will show that there are numerous initial Lagrangians leading to the observable physics, and the low-energy physics depends not only on the initial parameters, but also on the properties of compact spaces in a particular cascade. Therefore a variation of the initial parameters may be compensated by an appropriate variation of the properties of the cascade, leaving the low-energy physics almost unchanged. This diminishes the importance of a search for the "unique" Lagrangian of the Theory of Everything.

We also discuss the relationships between the parameters of the initial Lagrangian and the fundamental constants $\hbar$ and G determined in low-energy experiments.

11.2.3 Reduction cascades

Let us discuss the main idea in detail. Consider a D-dimensional space $\mathbb{M}_D$. Due to its metric fluctuations, new spatial regions with various geometries are continually born within it. We will be interested in regions with a direct product geometry of the form

$$\mathbb{U}_1 = T \times \mathbb{M}_{d_1} \times \mathbb{M}_{D_1}, \quad \mathbb{U}_1 \subset \mathbb{M}_D. \tag{11.4}$$

Here T is the timelike direction, $\mathbb{M}_{d_1}$ is a compact space of d_1 dimensions, and $\mathbb{M}_{D_1}$ is the main space whose metric fluctuations are studied at the next step of the cascade. We emphasize that the geometry changes not in the whole space $\mathbb{M}_D$ but only in its small region $\mathbb{U}_1$. All physical processes are considered from the viewpoint of an observer located inside the subspace $\mathbb{M}_{D_1}$.

Let us restrict our consideration to quantum fluctuations satisfying the following relation for Ricci scalars in the subspaces $\mathbb{M}_{d_1}$ and $\mathbb{M}_{D_1}$:

$$R_{d_1} \gg R_{D_1}. \tag{11.5}$$

For the simplest geometries, the larger the curvature of space, the smaller its volume. This is the type of geometries we will be studying. But the smaller the system volume, the faster the relaxation processes in it (see a discussion in [358]). Therefore, due to the condition (11.5), the processes in $\mathbb{M}_{d_1}$ advance much faster than those in $\mathbb{M}_{D_1}$. We will discuss the conditions needed to stabilize the volume of $\mathbb{M}_{d_1}$ and its geometry. The theory then becomes effectively D_1-dimensional, with d_1 compact extra dimensions. The parameters of the initial D-dimensional Lagrangian are renormalized, with their new values depending on the properties of the compact extra space $\mathbb{M}_{d_1}$.

What is described above is the first step of the cascade. The following steps are similar: due to the metric fluctuations in some volume of the newly formed space M_{D_1}, there arises a new geometry

$$\mathbb{U}_2 = T \times \mathbb{M}_{d_2} \times \mathbb{M}_{D_2}, \quad \mathbb{U}_2 \subset T \times \mathbb{M}_{D_1}, \tag{11.6}$$

resulting in segregation of another compact space $\mathbb{M}_{d_2}$. We choose those spaces among many for which the relation

$$R_{d_2} \gg R_{D_2}, \tag{11.7}$$

is satisfied similar to (11.5). The parameters of the reduced Lagrangian are renormalized once more; their values become dependent on the properties of the compact extra space $\mathbb{M}_{d_2}$ as well.

The succeeding steps form a cascade:

$$\begin{aligned} &\mathbb{M}_{D_1} \to \mathbb{M}_{D_2} \times \mathbb{M}_{d_2}; \\ &\mathbb{M}_{D_2} \to \mathbb{M}_{D_3} \times \mathbb{M}_{d_3} \to \cdots \to \mathbb{M}_3 \times \mathbb{M}_{d_{\rm final}}. \end{aligned} \tag{11.8}$$

Cascades differ from each other by the properties of their compact subspaces: their volume, topology, and geometry.

Each reduction consists of two stages: the first is the quantum formation of space of the form (11.4); the second is the classical evolution of this space, which results in stabilization of the compact extra space $\mathbb{M}_{d_i}$.

11.2.4 A step of the cascade in detail

Quantum formation of space

In this section we discuss the probability of quantum formation of geometries we are interested in. Consider a D-dimensional space. Quantum fluctuations in its small regions create subspaces of the form $\mathbb{M}_{d_1} \times \mathbb{M}_{D_1}$.

The absence of the Planck constant in the exponential function, see (10.140), may be thought of as a result of choosing the appropriate measurement units. It will be shown below that the Planck constant $\hbar$ naturally emerges simultaneously with the gravitational constant G after introduction of dimensional units. However, the statement on a unification of gravity and quantum theory would be premature. The essence of quantum mechanics is based on the summing rule for transition amplitudes (10.140), which is postulated originally.

What is the probability of such a process? The answer is far from clear even for the birth of a 4-dimensional space in the standard theory of gravity, linear in scalar curvature [196, 402]. Since we need only to verify that this probability is nonzero, let us approximate the Lagrangian by a linear theory. This is ensured by the condition (11.5).

The action in the space $\mathbb{M}_D$ is chosen in the form (10.145),

$$\begin{aligned} S_D &= N_0 \int d^D X \sqrt{|G^{(D)}|} F(R; a_n), \\ F(R; a_n) &= \sum_n a_n R^n, \quad a_1 = 1, \end{aligned} \tag{11.9}$$

where $G^{(D)} \equiv \det(G_{AB})$, R is the Ricci scalar, N_0 and $\{a_i\}$ are constants. Let us refer to S_D as the first generation action. The standard form is $F(R; a_n) = R - 2\Lambda$, i.e., $a_1 = 1$, $a_0 = -2\Lambda$.

The first step of a cascade begins in a region of small volume by forming a subspace (11.4) with a metric of the form

$$ds^2 = G_{AB} dX^A dX^B = \mathrm{N} dt^2 - g_{ab}(x) dx^a dx^b - \mathrm{e}^{2\beta(x)} \gamma_{ij}(y) dy^i dy^j, \tag{11.10}$$

where $g_{ab}(x)$ is the spatial part of the metric in $\mathbb{M}_{D_1}$, $\gamma_{ij}(y)$ is a positive-definite metric of the extra space $\mathbb{M}_{d_1}$ and $\mathrm{e}^{2\beta(x)}$ is a scaling factor (see [77, 101]).

Let us find the probability of producing such a space, approximating it by a linear dependence on R. Recall that owing to the form of the chosen metric (11.10), the relations (10.150) are valid.

As shown in [359], an extra space with an arbitrary geometry evolves into a space with a maximum number of Killing vectors for a given topology. So we can choose a maximally symmetric space with a constant Ricci scalar R_{d_1} and a curvature $k = \pm 1$.

The volume V_{d_1} of an internal space of unit curvature depends on its geometry since it is expressed in terms of its internal metric:

$$V_{d_1} = \int d^{d_1} y \sqrt{\left|G^{(d_1)}\right|}. \tag{11.11}$$

Let us emphasize that it sufficies to demand that only a small fraction of the volume of the initial space $\mathbb{M}_D$ has the geometry (11.10). The following discussion is concerned with the point of view of an internal observer in $\mathbb{M}_{D_1}$, who cannot get any information from outside.

We will use the slow-change approximation suggested in [83]:

$$|R_{d_1}| \gg |R_{D_1}|, \quad |R_{d_1}| \gg |f_{\rm der}|. \tag{11.12}$$

Substitution of Eq. (10.150) into Eq. (11.9) yields

$$\begin{aligned} F(R_D) &= F(R_{D_1} + R_{d_1} + f_{\rm der}) \\ &\simeq F(R_{d_1}) + F'(R_{d_1}) R_{D_1} + F'(R_{d_1}) f_{\rm der}. \end{aligned} \tag{11.13}$$

Also, the determinants satisfy the relation

$$|G^{(D)}| = \mathrm{e}^{2\beta(x)} \cdot |G^{(D_1)}| \cdot |G^{(d_1)}|. \tag{11.14}$$

Substituting these expressions into the action (11.9) and carrying out certain computations given explicitly in [83], we obtain a Lagrangian of a scalar-tensor theory of gravity in D_1 dimensions:

$$S_{D_1} = N_0' \int d^{D_1} x \sqrt{|G^{(D_1)}|}\, (\operatorname{sign} F')[R_{D_1} + \frac{K}{2}(\partial\phi)^2 - U], \tag{11.15}$$

$$K(\phi) = 3\left(\frac{F(\phi; a_n)''}{F(\phi; a_n)'}\right)^2 - \frac{d_1}{\phi}\frac{F(\phi; a_n)''}{F(\phi; a_n)'} + \frac{d_1}{4\phi^2}(d_1 + 2), \tag{11.16}$$

$$U(\phi) = -(\operatorname{sign} F(\phi; a_n)')\left[\frac{|\phi|}{d_1(d_1 - 1)}\right]^{d_1/2} \frac{F(\phi; a_n)}{F'(\phi; a_n)^2}. \tag{11.17}$$

In accordance with (11.12), we have kept only terms linear in the Ricci scalar.

The quantum birth of the Universe in linear theory has been studied by many authors. It is usually examined in the minisuperspace framework, where the interval is written in the form (11.20). The probability of quantum birth of a D_1-dimensional space was calculated in [99]. A lot of papers devoted to creation of universes in the presence of a scalar field has also appeared.

In our case, the situation is complicated by the fact that we consider nonlinear gravity and, in addition, there are extra dimensions whose stability should also be taken into account. The birth of an n-dimensional space with extra dimensions in the framework of standard gravity was considered in [103, 317], where stability regions of a compact subspace were also studied. The possibility of inflation in the presence of extra dimensions was explored in [103, 190]. Quadratic (in the Ricci scalar) gravity was investigated in this aspect in [388].

The Universe birth probabilities calculated in different approaches differ radically from one another [403]. This may be indicative of both an imperfection of modern theories and the complexity of the subject. The ultimate goal of such calculations is to determine the probability of appearance of a universe like ours. It would be unreasonable to expect this probability to be high, given that the parameters of the Universe are fine-tuned. In this case, calculating the probability is of purely academic interest, because there are no causal relationships between the universes. At present, to justify the promising study, it is probably necessary and sufficient to prove that the fraction of the universes like ours is nonzero in the framework of a specific approach. In our case, this means that the probability of each transition (11.8) in the cascade is nonzero.

The classical trajectories on which the action is stationary make a major contribution to the transition amplitude (10.140). Their shape depends on the boundary conditions and, in particular, on the properties of the manifold Σ_f. In our case, the metric on the hypersurface Σ_f is determined by the conditions (10.147), (10.148), (10.149). Therefore, we will seek classical trajectories subject to the same conditions on any section Σ between the sections $\Sigma_{\rm in}$ and Σ_f. The initial hypersurface $\Sigma_{\rm in}$ can either be absent altogether (the Hartle–Hawking approach) or have a "zero geometry" (the interval between any two points of this hypersurface is zero in Vilenkin's approach). We will show that the transition probability weakly depends on the properties of the hypersurface $\Sigma_{\rm in}$ in these cases. As an example, let us consider the formation probability of the structure

$$\Sigma_f = \mathbb{M}_3 \times \mathbb{M}_{d_{\rm final}}, \tag{11.18}$$

that emerges at the last step of the cascade. Classical trajectories make a major contribution to the transition amplitude. A topology change in the case of classical motion is unlikely. Therefore, classical trajectories consisting of hypersurfaces that also satisfy the condition (11.18) will be of importance for us. The topology of the D-dimensional Riemannian space between the sections Σ_{in} and Σ_f is then

$$\mathbb{R} \times \mathbb{M}_3 \times \mathbb{M}_{d_{\text{final}}}. \tag{11.19}$$

As above (see (10.148)), the inequality $R_3 \ll R_{d_{\text{final}}}$, which allows the results of the previous section to be used, is assumed. Indeed, in the approximation used, the action (11.29) transforms into a theory of the form (10.155) and then into the conventional Einstein–Hilbert action (10.165).

The action (10.165) has been repeatedly used to study the quantum birth of the Universe (see, e.g., [153, 196, 271, 402, 403, 421]). However, in such papers the presence of a scalar field is usually postulated, while in our approach this field constitutes the metric tensor components of the extra space. Therefore, we can use the results of numerous studies by briefly reproducing their main results. The quantum birth of the Universe is generally studied within the framework of minisuperspace in which the interval is written as [403]

$$\begin{aligned} ds^2 &= \sigma^2 \left[N(t)^2 dt^2 - a(t)^2 d\Omega_3^2\right], \\ \sigma^2 &= \frac{1}{12\pi^2 M_{Pl}^2}, \end{aligned} \tag{11.20}$$

where $N(t)$ is the lapse function and $a(t)$ is the scale factor. The wave function $\psi(a)$ satisfies the Wheeler–DeWitt equation

$$\left[\frac{\partial^2}{\partial a^2} - W(a)\right]\psi(a) = 0, \tag{11.21}$$

with the potential

$$W(a) = a^2(1 - H^2 a^2),\ a > 0, \quad H = \frac{\sqrt{U(\chi)}}{6\pi M_{\text{Pl}}^2}.$$

The birth of the Universe is described as a tunnelling transition with the forbidden region

$$0 < a < H^{-1}. \tag{11.22}$$

The wave function in this region is [403]

$$\psi(a) \simeq \exp\left[\int_a^{H^{-1}} \sqrt{-2W(a')}da'\right]. \tag{11.23}$$

The integral in this expression is ill-defined at the lower limit, where $a \to 0$. The approximation $R_3 \ll R_{d_{\rm final}}$ does not work in this region since $R_3 = k/a^2 \to \infty$, and no explicit expression is defined for the potential. The same problem also takes place in other models of a quantum birth of the Universe [295]. Nevertheless, the integral calculated in this way is meaningful in the limit

$$H \ll M_{\rm Pl}, \tag{11.24}$$

if the region $a \sim 0$ is small as compared to the entire integration domain. Since our Universe was formed at $H \sim 10^{-6} M_{\rm Pl}$, the inequality (11.24) holds even in the inflationary phase. The conclusion that the result weakly depends on the behaviour of the function near the singularity is also confirmed in papers by other authors. Thus, for example, an initial wave function of the form $\delta(a - a_{\rm in})$ was suggested in [388], while the decay of a metastable vacuum from a state with fixed energy was studied in [153, 421]. In both cases, the initial conditions were shown to weakly affect the transition probability. The quantum birth of universes in multidimensional gravity is discussed in detail in [103, 157].

In Vilenkin's approach, the probability of the birth of the Universe is $dP \propto \exp[+2/(3U(\chi))]$, while the Hartle–Hawking approximation yields $dP \propto \exp[-2/(3U(\chi))]$. Since the scalar field χ is uniquely related to the size of the extra space, the probability of birth of extra dimensions depends on their linear size. For all their differences, the main thing in both approaches is that the probability of the event is nonzero and hence the fraction of universes with given properties produced by a cascade of reductions is nonzero.

The classical evolution stage

To estimate the probability of the quantum birth of a space consisting of two subspaces, the main one and the extra one, we restricted the discussion in the previous Sec. 11.2.4 to an approximation linear in scalar curvature.

Let us confine ourselves to those initial parameters which lead to potentials having minima. Our numerical calculations indicate that such parameters do exist.

After nucleation, classical dynamics of these subspaces implies that the compact extra subspace evolves so that the field ϕ approaches the value ϕ_m, corresponding to a minimum of the potential $U(\phi)$.

As follows from the definition of the field ϕ (10.150) and the expression for the interval (11.10), the characteristic size of the space $\mathbb{M}_{d_1}$ is proportional to $e^{\beta} \propto 1/\sqrt{\phi}$. Therefore, the size of a compact *extra* space quickly stabilizes when the field reaches the value $\phi = \phi_m$. This corresponds to $R_{d_1} \to \text{const}$, $f_{\text{der}} \to 0$. Thus in the following discussion we can use the conditions

$$R_{d_1} = \text{const}, \quad f_{\text{der}} = 0, \tag{11.25}$$

which greatly simplify the calculations.

The equation for the scaling factor of the main subspace during a de Sitter stage has the form [101]

$$\frac{D_1(D_1-1)}{2}\left(\frac{\dot{a}}{a}\right)^2 = \Lambda - \frac{D_1(D_1-1)}{2}k. \tag{11.26}$$

We assume that the size of an extra space $\mathbb{M}_{d_1}$ has stabilized and $U(\phi) \simeq U(\phi_m) \equiv \Lambda$. Consequently, the scaling factor depends on time as

$$a(t) \propto e^{Ht}, \quad H = \frac{2\Lambda}{D_1(D_1-1)},$$

at large t. The size of the main subspace rapidly increases.

Thus, we have a D_1-dimensional quickly expanding space and a d_1-dimensional compact extra space. The curvature-linear approximation is sufficient to obtain this result, but to advance further we will need a more accurate expression for the reduced action. The latter could be derived using the conditions (11.25). Expanding the relation $F(R; a_n) = F(R_{d_1} + R_{D_1}; a_n)$ into a Taylor series and integrating over the extra-dimensional coordinates, we obtain

$$S_{D_1} = N_0' \int d^{D_1}X \sqrt{|G^{(D_1)}|} F(R_{D_1}; \tilde{a}_n). \tag{11.27}$$

Here the new parameters $\tilde{a}_n$ are functions of R_{d_1}, $n > 0$.

We arrive at the second generation of the action (11.27), which is similar to the first generation of the action (11.9) with changed numerical values of the parameters $\tilde{a}_n$. The dimensionality of the main space has been reduced, $D_1 < D$, and a new compact extra space of $d_1 = D - D_1$ dimensions has been formed.

The size of the space $\mathbb{M}_{D_1}$ is much larger than the size of $\mathbb{M}_{d_1}$. In this discussion we are concerned only with quantum fluctuations creating spaces satisfying such a relation. The subsequent dynamics further increases their disparity.

The second step of a cascade is analogous to the first one with substitution $\mathbb{S}_D \to \mathbb{S}_{D_1}$: in a small region of the space $\mathbb{M}_{D_1}$ there occurs a quantum fluctuation which creates a subspace with the topology $\mathbb{M}_{D_2} \times \mathbb{M}_{d_2}; D_2 + d_2 = D_1$. As a result of classical dynamics, the size of the space $\mathbb{M}_{d_2}$ is stabilized while the space $\mathbb{M}_{D_2}$ expands.

If we do not want to be concerned with excitations of the compact space $\mathbb{M}_{d_1}$, we should only consider such quantum fluctuations in $\mathbb{M}_{D_1}$ that satisfy $R_{d_2} \ll R_{d_1}$.

11.3 Quadratic gravity as an explicit example

In this section we will be guided by the formulas derived in Chapter 9.4.5. The initial gravitational field action includes all powers of the Ricci scalar and other invariants. A vast majority of the works on the subject uses some finite polynomial in the Ricci scalar. The choice of a particular polynomial may be justified as follows. Consider a quantum fluctuation that produces a geometry with a characteristic value R_0 of the scalar curvature. Then the initial Lagrangian (11.9) may be approximated by a finite polynomial:

$$F(R, a_n) \simeq \sum_{k=-K}^{K'} b(R_0)_k (R - R_0)^k. \tag{11.28}$$

The specific values of K and K' are chosen according to the author's purposes. The coefficients $b_k(R_0)$ depend on the location of the expansion (11.28) and vary in a wide range.

Consider quadratic gravity in the space $\mathbb{M}_D$ with the action of the form

$$S_D = \frac{N_0}{2} \int d^D X \sqrt{|G^{(D)}|} \left[R_D(G_{AB}) + C R_D^2(G_{AB}) \right.$$
$$\left. + C_1 R_{AB} R^{AB} + C_2 \mathcal{K} - 2\Lambda \right] + \int_{\partial M_D} K d^{D-1} \Sigma, \tag{11.29}$$

where we have also included the Ricci tensor squared and the Kretschmann scalar $\mathcal{K} = R_{ABCD} R^{ABCD}$. The boundary term $(\partial \mathbb{M}_D)$ introduced by Hawking and Gibbons does not affect the classical dynamics and can be ignored in what follows [99].

Let us try to find the values of the parameters that allow the formation of a universe similar to ours. In this case a set of parameters of the Lagrangian $\{a_n\}$ (see (11.9)) is $\{C, C_1, C_2, \Lambda\}$.

Following the steps outlined above, we will find the form of the action (11.29) reduced to the space $\mathbb{M}_{D_1}$. An action pertaining to only $\mathbb{M}_{D_1}$ can be recovered by integrating the action in $\mathbb{M}_D$ (11.29) over $\mathbb{M}_{d_1}$.

The Ricci scalar can be expressed using Eqs. (10.150):

$$R_D(G_{AB}) = R_{D_1}(g_{ab}) + \phi(x), \tag{11.30}$$

where we have set $f_{\rm der} = 0$, see (11.25). The decompositions of $R_{AB}R^{AB}$ and $\mathcal{K}$ are given by (see [83])

$$\begin{aligned} R_{AB}R^{AB} &= R_{ab}R^{ab} + \mathrm{e}^{-4\beta(x)} \cdot R_{\mu\nu}R^{\mu\nu}, \\ \mathcal{K} &= \mathcal{K}(g_{ab}) + \mathrm{e}^{-4\beta(x)} \cdot \mathcal{K}(\gamma_{ij}), \end{aligned} \tag{11.31}$$

where variables with the indices A, B correspond to the metric G_{AB}, those with the indices a, b to g_{ab}, and with μ, ν to the metric γ_{ij}, (i.e., they correspond to the spaces $\mathbb{M}_D$, $\mathbb{M}_{D_1}$ and $\mathbb{M}_{d_1}$, respectively).

To advance further, recall that we are considering the d_1-dimensional metric γ_{ij} of a constant curvature k, so that we can express the Riemann tensor, the Ricci tensor, and the Ricci scalar in $\mathbb{M}_{d_1}$ space in terms of its curvature:

$$R^{\mu\nu}{}_{\rho\eta} = k\,\delta^{\mu\nu}{}_{\rho\eta}, \quad R^{\nu}_{\mu} = k\,(d_1 - 1)\delta^{\nu}_{\mu}, \quad R'_{d_1} \equiv k\,d_1(d_1 - 1), \tag{11.32}$$

where $\delta^{\mu\nu}{}_{\rho\eta} = \delta^{\mu}_{\rho}\delta^{\nu}_{\eta} - \delta^{\mu}_{\eta}\delta^{\nu}_{\rho}$. R'_{d_1} represents the characteristic curvature scale of the extra dimensions. The expressions for the squared Riemann and Ricci tensors are derived from (11.32); the former is by definition the Kretschmann scalar:

$$\begin{aligned} R_{\mu\nu}R^{\mu\nu} &= d_1[k(d_1 - 1)]^2, \\ \mathcal{K}(\gamma_{ij}) &\equiv R_{\mu\nu\rho\eta}R^{\mu\nu\rho\eta} = 2d_1(d_1 - 1)k^2. \end{aligned} \tag{11.33}$$

Now we can rewrite (11.31) substituting $\mathrm{e}^{\beta(x)}$ from (10.150) and using (11.33):

$$\begin{aligned} R_{AB}R^{AB} &= R_{ab}R^{ab} + \frac{1}{d_1}\phi(x), \\ \mathcal{K} &= \mathcal{K}(g_{ab}) + \frac{2}{d_1(d_1 - 1)}\phi(x). \end{aligned} \tag{11.34}$$

After plugging (11.30) and (11.34) into the action (11.29) and grouping the terms we obtain

$$\begin{aligned} S_D = \frac{N_0}{2} \int d^D x \sqrt{|G^{(D)}|} \Big\{ & R_{D_1}(g_{ab})(1 + 2C\phi) + C R^2_{D_1}(g_{ab}) - 2\Lambda \\ & + C_1 R_{ab} R^{ab} + C_2 \mathcal{K}(g_{ab}) + \phi + \left(C + \frac{C_1}{d_1} + \frac{2C_2}{d_1(d_1 - 1)} \right) \phi^2 \Big\} \\ = \frac{N_0}{2} \int d^D x \sqrt{|G^{(D)}|} \cdot \mathcal{L}(g_{ab}). & \end{aligned} \tag{11.35}$$

The expression in brackets has been denoted $\mathcal{L}(g_{ab})$ for convenience; it does not depend on the coordinates of the extra space $\mathbb{M}_{d_1}$.

To find the action in $\mathbb{M}_{D_1}$ we have to integrate (11.35) over the space $\mathbb{M}_{d_1}$ using the appropriate volume definition (11.11). Substitution of Eq. (11.14) for $\sqrt{|G^{(D)}|}$, Eq. (10.150) and (11.32) for $e^{2\beta(x)}$ yields:

$$\begin{aligned} S_D &= \frac{N_0}{2} \int d^D x \sqrt{|G^{(D)}|} \cdot \mathcal{L}(g_{ab}) \\ &= \frac{N_0}{2} \int d^{D_1} x \; e^{2\beta(x) d_1} \sqrt{|G^{(D_1)}|} \cdot \mathcal{L}(g_{ab}) \int d^{d_1} x \sqrt{|G^{(d_1)}|} \\ &= \frac{N_0 V_{d_1}}{2} \int d^{D_1} x \sqrt{|G^{(D_1)}|} \left(\frac{R'_{d_1}(G^{(d_1)})}{\phi(x)} \right)^{d_1/2} \cdot \mathcal{L}(g_{ab}). \end{aligned} \tag{11.36}$$

Suppose that there exists a minimum of the potential $U(\phi)$. The field $\phi(x)$ rapidly relaxes to it and stays fixed during the low-energy processes (see [359] for a discussion). This case is the most natural since the relaxation time is proportional to the scale of the extra space $\mathbb{M}_{d_1}$, which is small as compared to the scale of the space $\mathbb{M}_{D_1}$.

Assuming that these conditions are satisfied, let us make a conformal transformation of the form (see, e.g., [60])

$$\begin{aligned} & g_{ab} = |f(\phi_m)|^{-2/(D_1-2)} \tilde{g}_{ab}, \\ & f(\phi) \equiv \phi^{-d_1/2}(x)[1 + 2C\phi(x)], \\ & R_{D_1} = |f(\phi_m)|^{2/(D_1-2)} \tilde{R}_{D_1}, \\ & R_{ab} R^{ab} = |f(\phi_m)|^{4/(D_1-2)} \tilde{R}_{ab} \tilde{R}^{ab}, \\ & \mathcal{K} = |f(\phi_m)|^{8/(D_1-2)} \tilde{\mathcal{K}}, \\ & \sqrt{|G^{(D_1)}|} = |f(\phi_m)|^{-D_1/(D_1-2)} \sqrt{|\tilde{G}^{(D_1)}|}, \end{aligned} \tag{11.37}$$

which, being applied to Eq. (11.36), brings us to the initial form of the action (compare with (11.29)):

$$S_{D_1} = \frac{N_0^1}{2}\int d^{D_1}x\sqrt{|G^{(D_1)}|}\Big\{R_{D_1}(g_{ab}) + C^{(D_1)}R_{D_1}(g_{ab})^2 + C_1^{(D_1)}R_{ab}R^{ab}(g_{ab}) + C_2^{(D_1)}\mathcal{K}(g_{ab}) - 2\Lambda^{(D_1)}\Big\}, \qquad (11.38)$$

where $N_0^1 = N_0 V_{d_1}$ and the tildes were omitted for short. The new parameters are expressed in terms of the old ones as follows:

$$C^{(D_1)} = \mathrm{sign}(f(\phi_m))\,|f(\phi_m)|^{(4-D_1)/(D_1-2)}\phi_m^{-d_1/2}C, \qquad (11.39)$$

$$\Lambda^{(D_1)} = \mathrm{sign}(f(\phi_m))|f(\phi_m)|^{-D_1/(D_1-2)}\phi_m^{-d_1/2} \times\left[\Lambda - 1/2\left(\phi_m + \left(C + \frac{C_1}{d_1} + \frac{2C_2}{d_1(d_1-1)}\right)\phi_m^2\right)\right], \qquad (11.40)$$

$$C_1^{(D_1)} = \mathrm{sign}(f(\phi_m))\,|f(\phi_m)|^{(4-D_1)/(D_1-2)}\phi_m^{-d_1/2}C_1, \qquad (11.41)$$

$$C_2^{(D_1)} = \mathrm{sign}(f(\phi_m))\;|f(\phi_m)|^{(8-D_1)/(D_1-2)}\phi_m^{-d_1/2}C_2. \qquad (11.42)$$

where $f(\phi)$ is defined in (11.37). Recall that we are considering the case where the field ϕ is already at its minimum, $\phi = \phi_m$, so the kinetic terms are neglected. Eqs. (11.39)–(11.42) connect the old and new parameters after a single reduction.

Thus a single step of reduction to a space of smaller dimension only changes the numerical values of the initial parameters. Meanwhile, the form of the action remains invariable.

The next reduction leads to similar relations for the parameters $C^{(D_2)}$, $C_1^{(D_2)}$, $C_2^{(D_2)}$, $\Lambda^{(D_2)}$ with the substitutions $C \to C^{(D_1)}$, $C_1 \to C_1^{(D_1)}$, $C_2 \to C_2^{(D_1)}$, $\Lambda \to \Lambda^{(D_1)}$. Thus we have obtained recurrence formulas for the parameters.

Again, the action (11.38) for a subspace $\mathbb{M}_{D_1}$ coincides in its form with the initial action (11.29) for $\mathbb{M}_D$, but with renormalized parameters $C^{(D_1)}$, $C_1^{(D_1)}$, $C_2^{(D_1)}$, $\Lambda^{(D_1)}$.

11.3.1 Formation of low-energy physics

Here we use the results of Sec. 10.4.2.

It was shown in the previous section that the form of the action (11.29) does not change after the reductions (11.10). Now we will consider the final reduction of an arbitrary cascade of $(n+1)$ reductions, the reduction leading to four-dimensional space-time.

As was shown in [83], the action (11.29) may be reduced to the four-dimensional form

$$S \simeq \frac{N_0 V_d}{2} \int d^4x \sqrt{|G^{(4)}|}[R_4 + K(\phi)(\partial\phi)^2 - 2U(\phi)], \tag{11.43}$$

$$U(\phi) = \frac{-\operatorname{sign}(1 + 2C^{(D_n)}\phi)}{2} \cdot \frac{C_{\text{tot}}^{(D_n)}\phi^2 + \phi - 2\Lambda}{(1 + 2C^{(D_n)}\phi)^2} \left[\frac{|\phi|}{d_1(d_1 - 1)}\right]^{\frac{d_1}{2}}, \tag{11.44}$$

$$K(\phi) = \frac{(C^{(D_n)})^2\phi^2(d_1^2 - 2d_1 + 12) + C^{(D_n)}d_1^2\phi + \frac{1}{4}d_1(d_1 + 2)}{(1 + 2C^{(D_n)}\phi)^2\phi^2} + \frac{C_1^{(D_n)} + C_2^{(D_n)}}{2\phi(1 + 2C^{(D_n)}\phi)}, \tag{11.45}$$

where $C_{\text{tot}}^{(D_n)} \equiv C^{(D_n)} + \frac{C_1^{(D_n)}}{d_1} + \frac{2C_2^{(D_n)}}{d_1(d_1-1)}$ (compare with (11.16), (11.17), which are written for the case $C_1^{(D)} = C_2^{(D)} = 0$). These equations, though derived for the action (11.29), are still applicable after any number of reductions since the reductions do not change the form of the action, as was shown in the previous section.

In the vicinity of a minimum, $U(\phi_m) \equiv \min(U(\phi))$, the potential can be expanded in a Taylor series, so that the action (11.43) becomes

$$S \simeq \frac{N_0 V_d}{2} \int d^4x \sqrt{|G^{(4)}|} \Big[R_4 + K(\phi_m)(\partial\phi)^2 - 2U(\phi_m) - U''(\phi_m)(\phi - \phi_m)^2 - \frac{1}{3}U'''(\phi_m)(\phi - \phi_m)^3 - \cdots\Big]. \tag{11.46}$$

The measurement unit **I** is still arbitrary. Let the units be related by the expressions (10.160),

$$\mathbf{I} = \alpha \cdot \text{cm},$$

where α is a yet unknown parameter. We now turn on to the standard units of length (10.161):

$$S = N_0 V_d c\alpha^2 \int dt d^3x \Big[R_4 + \frac{1}{2}K(\phi_m)(\partial\phi)^2 - \alpha^2 U(\phi_m) - \alpha^2\frac{1}{2}U''(\phi_m)(\phi - \phi_m)^2 - \alpha^2\frac{1}{6}U'''(\phi_m)(\phi - \phi_m)^3\Big]. \tag{11.47}$$

Here, $x \to \alpha x$, $\sqrt{|G^{(4)}|} = c_I = \alpha c$, $R_4 \to R_4/\alpha^2$, and $(\partial\phi)^2 \to (\partial\phi)^2/\alpha^2$.

Since the expression (10.161) is the low-energy limit of the action (11.29), it should adequately describe purely gravitational phenomena. Moreover, the expression (10.161) makes it possible to explain the origin of the inflaton potential and the cosmological constant. The effective action (10.161) contains the initial parameters of the theory without involving fundamental constants, such as the Planck constant $\hbar$ and the gravitational constant G. Now we will find relations between those values.

Recall that

$$m_\Phi \equiv \alpha \sqrt{\frac{U''(\phi_m)}{K(\phi_m)}},$$

(see (10.164)) and, under the assumption $K(\phi_m) > 0$, we introduce the variable

$$\Phi = \sqrt{\frac{c^4}{16\pi G} K(\phi_m)}(\phi - \phi_m).$$

The action (10.161) in terms of Φ is

$$S = \frac{16\pi G}{c^4} N_0 V_d c \alpha^2 \int dt d^3x \left[\frac{c^4}{16\pi G} R_4 + \frac{1}{2}(\partial\Phi)^2 - \frac{c^4}{16\pi G}\alpha^2 U(\phi_m) \right.$$
$$\left. - \frac{1}{2} m_\Phi^2 \Phi^2 - \frac{1}{6}\sqrt{\frac{16\pi G}{c^4}} m_\Phi^2 \frac{U'''(\phi_m)}{U''(\phi_m)\sqrt{K(\phi_m)}} \Phi^3 \right]. \quad (11.48)$$

We arrive at the conventional form of the action for the scalar field Φ interacting with gravity, see (10.165):

$$S = \frac{1}{\hbar} \int dt d^3x \left(\frac{c^4}{16\pi G} R_4 + \frac{1}{2}(\partial\Phi)^2 - \Lambda - \frac{1}{2} m_\Phi^2 \Phi^2 + \lambda_3 \Phi^3 \right).$$

Comparing (10.165) with (11.48), we obtain the relations

$$\frac{1}{\hbar} = \frac{16\pi G}{c^3} N_0 V_d \alpha^2, \quad (11.49)$$

$$\Lambda = \frac{c^4}{16\pi G}\alpha^2 U(\phi_m), \quad (11.50)$$

$$\lambda_3 = -\frac{1}{6}\sqrt{\frac{16\pi G}{c^4}} m_\Phi^2 \frac{U'''(\phi_m)}{U''(\phi_m)\sqrt{K(\phi_m)}}. \quad (11.51)$$

Eliminating the parameter α from Eqs. (11.49), (11.50), (11.51) and (10.164), we obtain

$$\frac{U(\phi_m)K(\phi_m)}{U''(\phi_m)} = \frac{16\pi G}{c^4}\frac{\Lambda}{m_\Phi^2}, \tag{11.52}$$

$$N_0 V_d \frac{K(\phi_m)}{U''(\phi_m)} = \frac{c^3}{16\pi G\hbar m_\Phi^2}, \tag{11.53}$$

$$\frac{U'''(\phi_m)}{U''(\phi_m)\sqrt{K(\phi_m)}} = -6\sqrt{\frac{c^4}{16\pi G}}\frac{\lambda_3}{m_\Phi^2}. \tag{11.54}$$

The right-hand sides of (11.52)–(11.53) contain fundamental constants while the left-hand sides depend on the initial parameters. If the field Φ is associated with the inflaton, its mass is about 10^{13} GeV. The parameter λ_3 defined in (11.51) is not yet measured and may be considered as a prediction of our approach. In numerical calculations we kept in mind that $\lambda_3 < 10^{-12}M_{\rm Pl}$. It provides inflation which does not contradict the observational data.

As was discussed in Sec. 10.4.2, the conversion factor between the units of length $\alpha = \mathbf{I}/cm$ can also be found by comparison with the characteristic size of the extra space ${V_d}^{1/d}$ in the units of $\mathbf{I}$, $\alpha = L_d/{V_d}^{1/d}$. Here L_d is the size of the hypothetic extra space expressed in cm. The values of $L_d \leq 10^{-17}$ do not contradict the experimental data. Taking into account Eq. (10.164), we find the constraint on the initial parameters

$$L_d = m_\Phi V_d^{1/d}\sqrt{\frac{K(\phi_m)}{U''(\phi_m)}} < 10^{-17}. \tag{11.55}$$

The relations (11.52)–(11.55) allow us to relate the initial parameters of the theory N_0, C, C_1, C_2, Λ from the action (11.29) with the observed constants c, $\hbar$, G, Λ, m_Φ and to predict the value of inflaton coupling constant λ_3. The question is whether or not we can find the initial parameters satisfying the constraints (10.166) and (11.55). We have proved the existence of such parameters by numerical simulations of reduction cascades. As is shown in the next section, there are numerous sets of initial parameters satisfying those conditions.

11.3.2 Numerical computations

We generate the random initial parameter set $\{C, C_1, C_2, \Lambda\}$ and try to find the cascade that leads to the 4-dimensional space-time and satisfies all the constraints laid out in Sec. 11.3.1. The cascade is completely specified by the dimensions of the spaces that undergo reductions. For instance, we take a 14-dimensional space as our starting point and consider the reduction to an 11-dimensional one, which stabilizes when the field ϕ reaches its minimum. This new space is in turn reduced to an 8-dimensional one, and finally to the 4-dimensional one. Thus, we have a cascade that could be designated as $\langle 15 \to 11 \to 8 \to 4 \rangle$.

We have carried out numerical simulations to assess the density of the initial parameters $\{C, C_1, C_2, \Lambda\}$ that can lead to the observable 4-dimensional space-time. The greater this density, the less precisely one has to specify the set of initial parameters of the theory. In the event that every set of parameters $\{C, C_1, C_2, \Lambda\}$ leads to the observable 4D space-time, there is no need to specify the initial parameters at all, because all of them can lead to our space-time. If this could be done for the initial space-time of any dimensionality, then dimensionality also does not have to be specified.

These reduced space-times have their own sets of parameters $\{C, C_1, C_2, \Lambda\}$. After each reduction the parameters are transformed according to (11.39)–(11.42). Using these recurrence formulas we can obtain the last set of parameters before the final reduction. Plugged into (11.44) and (11.45) they give the potential and kinetic energies at the last step of reductions. Finally, if this parameter set satisfies the conditions (11.52)–(11.55), we can conclude that they correspond to the observable 4-dimensional space-time.

A set of initial parameters $\{C, C_1, C_2, \Lambda\}$ is represented by a point in the parametric hyperspace (C, C_1, C_2, Λ). We shall refer to those parameter sets that lead to observable constants (i.e. satisfy (11.52)–(11.55)) as the *solutions*. We are interested in the density of solutions in the parametric hyperspace. To assess the density visually we have to use its plane projections — for instance, projections onto planes (C_1, C_2) or (C, Λ). Figure 11.1 presents the results of numerical computation — projections of a four-dimensional solution image onto planes (C, C_1) and (C, Λ).

It can be seen that there are numerous solutions. This means that the observed low-energy physics is reproduced by a wide range of initial parameters of our theory.

Recall that we took into account only the simplest geometries of the hyperspaces undergoing reductions — only the maximally symmetrical ones. Including other geometries into consideration would yield additional solutions. If these solutions form a continuum in the parametric

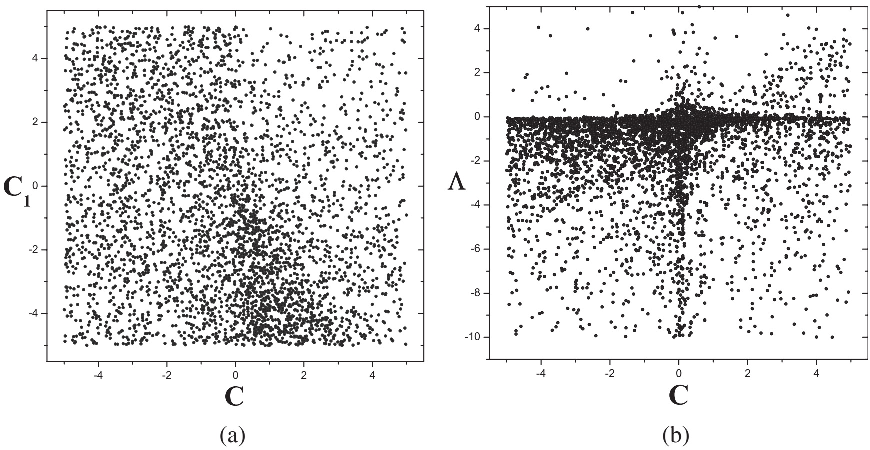


Figure 11.1 Results of numerical computation: plane projections of the points in the hyperspace (C, C_1, C_2, Λ) corresponding to solutions: (a) a projection onto the plane (C, C_1) (b) a projection onto the plane (C, Λ).

space, any point in that space (i.e. any set of initial parameters) would lead to the observable physics.

11.4 Discussion

The prospective theory, called a "Theory of Everything" (TOE), should solve the problem of fine tuning in the Universe. How are the initial parameters of TOE chosen and why do they allow very complex structures to arise? A way to approach this problem was found by proposing the existence of numerous vacua with different properties — a "landscape" in the modern terminology. The landscape derived in the scope of string theory is an important advance, but the questions to be answered still remain. Is the concept of "strings" indispensable or will the assumption of multiple dimensions alone suffice? How crucial are the values of the initial parameters? What is the probability of getting to a particular vacuum? Which additional values besides the metric do we need to include?

In the framework of the theory described in this chapter, these questions can be answered as follows:

— For a low-energy "landscape" to exist, we only need to assume the existence of multiple dimensions. The rest of the string theory tools as well as incorporation of additional fields are unnecessary. The cascade reduction mechanism effectively produces the low-energy physics with various parameters.

— Changes in the initial parameters do not substantially affect the probability of universe formation.

— The values of the Plank constant and the gravitational constant, being expressed in the same units I — see discussion in Sec. 10.4.2 — differ in different universes and depend on the choice of a reduction cascade for a given set of initial parameters. Those constants change with time during the early stages of the universe formation.

It has been shown that there are numerous values of initial parameters of the theory which can be "connected by a cascade" with observed fundamental constants. Particular numerical values of the initial parameters are therefore not as important as was previously thought.

The "landscape" concept implies that numerous low-energy theories with various properties originate from a theory with unique initial parameters. Here we are concerned with the opposite scenario that the observed physics is derived from numerous initial theories. This "inverse landscape" model puts to doubt the significance of a search for the unique Lagrangian of a TOE.

A set of all such values is rather large although it does not constitute a continuum in the parametric hyperspace. This may be attributed to the limited subset of all possible cascades that we have studied — only those consisting of absolutely symmetric compact spaces. Evidently, an extension of this subset to include all possible geometries will increase the number of acceptable parameters.

Conjecture

For any given set of initial parameters, there exists a cascade of reductions leading from a multidimensional space to a universe of our type.

The credibility of this hypothesis should be fully examined in future research.

We have already studied the conditions making it possible to create universes with different properties. Extra dimensions and quantum fluctuations are necessary attributes. Let us pose a question: are extra dimensions really so necessary in the present case?

We will show that already the existence of quantum fluctuations is sufficient for making certain inferences.

Random potential

When creating a theory, one usually postulates some specific form of the Lagrangian. The interaction constants are assumed to be small, so that quantum corrections to the originally chosen Lagrangian are also small. In the present case, it is important that quantum corrections lead to emergence of an infinite number of additional terms in an initially simple Lagrangian. The corrections are really small for weak fields, but for strong fields it is far from being so.

On renormalizability

A widespread viewpoit is that there are "good", renormalizable theories and "bad", nonrenormalizable ones. Briefly, a theory is called renormalizable if quantum corrections do not lead to a necessity of inserting an infinite number of new terms into the initial Lagrangian. Strictly speaking, **ALL theories are nonrenormalizable.** Indeed, if one recalls that all fields are supposed to be coupled to gravity, which is a nonrenormalizable theory, this statement becomes almost trivial. From this viewpoint, the only "good" — renormalizable theory is the one containing all possible terms.

As an example, consider the Lagrangian of a scalar field with self-interaction

$$L = \frac{1}{2}(\partial_\mu \varphi)^2 - \frac{m^2}{2}\varphi^2 - \frac{\lambda}{4}\varphi^4. \tag{11.56}$$

One-loop quantum corrections to the potential have the form [272]

$$\delta V = \frac{(3\lambda\varphi^2 + m^2)^2}{64\pi^2} \ln \frac{(3\lambda\varphi^2 + m^2)}{2m^2} - a\varphi^2 - b\varphi^4. \tag{11.57}$$

The last two terms renormalize the mass and the coupling constant in the initial potential and depend on the way of renormalization. But the first term changes the very form of the potential. Many-loop corrections add new terms to the potential.

It is easy to see, comparing the expressions (11.56) and (11.57), that the new terms are small relative to the initial ones in the range $\varphi \ll m \cdot \exp(1/\lambda)$. If one chooses for estimates the values of $m = 100$ GeV, $\lambda = 0.1$ (we mean here that the field is not an inflaton), then the quantum corrections to the potential become large at $\varphi \sim 10^6$ GeV. It is quite a large value if we study the interactions with accelerators, but at the early inflationary stage of our Universe's evolution the fields reached values at the level $\varphi > 10^{19}$ GeV, and at chosen values of the parameters it is necessary to take into account the whole infinite series of additional terms in the Lagrangian (11.56). Even more than that, restrictions on the scalar field magnitude are still stronger. Indeed, the logarithm in the expression (11.57) is a result of the summing of an infinite series that converges only at $\varphi < m/\sqrt{3\lambda}$. Besides, even from the form of the Lagrangian (11.56) it is seen that the interaction potential becomes comparable with the mass term at $\varphi \sim m\sqrt{2/\lambda}$. The latter two estimates agree well with each other and give a much smaller field magnitude at which one could call the quantum corrections small.

Thus when considering strong-field phenomena, such that $\varphi > m/\sqrt{\lambda}$, it is necessary to include all additional terms that inevitably emerge when taking into account the quantum corrections. The shape of the potential becomes much more complicated than could be imagined on the basis of the low-energy limit of the theory. Thus, in a mountain country one can speak of smooth surfaces with a small curvature only in valleys, i.e., at minima of the potential energy, but having ascended to a certain height, it becomes evident that the relief is much more complex.

The interaction potential of a scalar field is usually postulated in the simplest form suitable for the purposes of a specific study. Then the renormalizability requirement is not compulsory because it is assumed that inclusion of gravitational effects at the Planck scale will lead to regularization of the integrals. In conventional field theories the fields are weak indeed, and quantum corrections are reduced to renormalizations of the Lagrangian parameters and an assumption that the finite corrections are small. From the above arguments it follows that at the birth of our Universe, at large field magnitudes, the quantum corrections were most probably of the same order as the basic terms, and the form of the Lagrangian was much more involved than could be assumed at the outset. Therefore, instead of postulating a simple form of the Lagrangian and trying to prove that it is valid at high energies as well, let us do just the opposite. Namely, let us accept as a fact that the potential is a polynomial containing all powers of the field φ [353, 355]. The contribution of each term into the renormalization is a result of a complex interference of contributions from interactions with all sorts of fields. This reasoning leads to the idea that the specific numerical values of different polynomial coefficients are weakly related to each other. We suggest to accept the following statement as a postulate:

The only correlation among the polynomial coe cients of the potential is the condition $|V(\varphi)| \le M_{\rm Pl}^4$. $(\star)$

The latter inequality allows one to avoid complications related to quantization of gravity at the Planck scale. We will show that this postulate leads to consequences of interest and is intrinsically noncontradictory. As a working example, consider a scalar field Lagrangian of the form

$$L = \frac{1}{2}(\partial_\mu \varphi)^2 - V(\varphi), \quad V(\varphi) = \sum_{n=0}^{\infty} a_n \frac{\varphi^n}{M_{\rm Pl}^{n-4}}. \tag{11.58}$$

Potentials with an arbitrary but finite polynomial order are used in the model of Λ inflation [280]. The field φ is defined in the range $(-\infty, +\infty)$. The factors at powers of φ in the polynomial $V(\varphi)$ are random numbers. The random number distribution law is irrelevant.

Figure 11.2 presents a characteristic plot of the potential in some range of the φ field.

At each minimum, the potential $V(\varphi)$ is approximated in a natural way: $V(\varphi) \approx V(\varphi_m) + a\phi^2 + b\phi^3 + c\phi^4$, $\phi = \varphi - \varphi_m$. Usually such a potential is postulated from the very beginning with certain constants a and b, and the constant a is connected with the mass of the φ field, $a = m_\varphi^2/2$, if $a > 0$. Our Universe is situated at one of such minima. Let us also note that the Lagrangian (11.57) is a special case of a more general Lagrangian which should contain quantum corrections to the kinetic term.

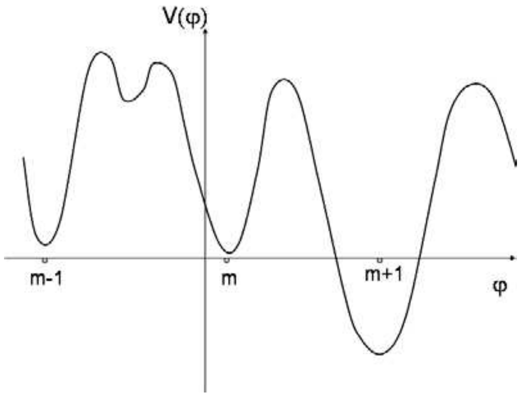


Figure 11.2 A possible shape of the effective potential.

In the next section, we discuss cosmological consequences of the introduced postulates.

Quantum fluctuations as a generator of universes

All (quasi-)stationary states are located at minima of the potential, and our Universe, not being an exception, is located in one of them. Since there is a countable number of minima, each being characterized by its own energy density, it seems unlikely that our Universe has gotten into a particular minimum with a small energy density. To estimate this probability, let us assume that the probability of finding a minimum of the potential with the energy density $\rho_V^{(m)} = V(\varphi_m)$ in the range $d\rho_V^{(m)}$ is

$$dP(\rho_V^{(m)}) = d\rho_V^{(m)}/M_{\text{Pl}}^4, \tag{11.59}$$

i.e., we assume a homogeneous distribution of $\rho_V^{(m)}$ in the whole range $(0,\, M_{\text{Pl}}^4)$. The expected vacuum energy density value in our Universe is $\rho_V \approx 10^{-123} M_{\text{Pl}}^4$. Consequently, the fraction of universes with vacuum energy densities like ours is $\sim 10^{-123}$. Still the number of such universes is countable because such is the number of all minima. Quantum fluctuations that emerge, in particular, near these minima, create a countable number of universes of our type in the course of inflation.

A causally connected region, having emerged, eventually splits into more and more causally disconnected regions, and some of them, due to field fluctuations, move to the closest maximum of the potential. Consider

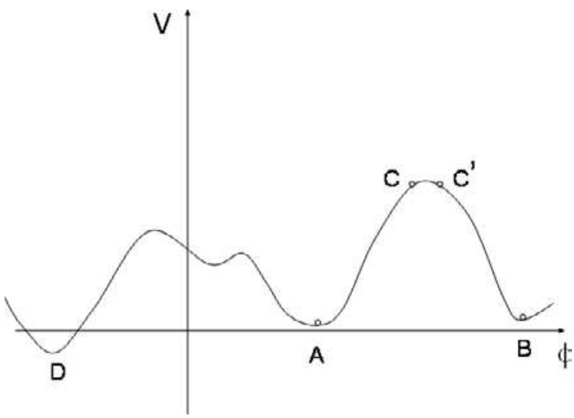


Figure 11.3 Fractal structure formation due to quantum fluctuations of a scalar field near a maximum of the potential.

a fluctuation close to a maximum (point C in Fig. 11.3). In a time equal to $1/H$, this spatial region will split into e^3 causally disconnected regions, each with its own field value inside. The mean value of the φ field in some of these regions will get into another side of the maximum (point C' in Fig. 11.3). Each of the regions will in turn, in a time of $1/H$, consist of e^3 causally disconnected regions, and some of them will pass through the maximum in the opposite direction. Such a process is self-reproducing, and already after a few steps there will be a fractal-type picture in the neighborhood of the maximum.

It is known that adjacent regions located at different sides of a maximum of the potential are separated by a wall [335]. Fields in these regions roll down to different (neighboring) minima, while the wall energy density grows. Thus neighboring universes turn out to be separated by field walls with large energy densities.

Thus from a consistent application of the postulate ($\star$) it follows that our Universe has formed from a single region bounded by a closed wall. The inflationary mechanism has provided an exponential growth of its size (from the viewpoint of an internal observer), so that the modern size of the visible part of the Universe of $\sim 10^{28}$ cm turns out to be many orders of magnitude smaller than the characteristic size of the closed region, $\sim 10^{10^{12}}$ cm [275]. Consequently, such walls are unobservable.

This mechanism also makes it possible to select universes according to other parameters, for instance, by particle masses. An interaction of a scalar field with fermions is usually postulated in the form of a Yukawa coupling,

$$V_F = g\varphi\bar{\psi}\psi. \tag{11.60}$$

In this case we are facing a serious difficulty. The point is that the suitable minimum of the potential may turn out to be quite far from the value $\varphi = 0$. Consequently, the contribution to the fermion mass $\mu_F = g\varphi_m$ due to interaction with the inflaton will be anomalously large. Even if a direct interaction of the form (11.60) is absent from the beginning, it emerges due to quantum corrections. The problem is solved if one takes into account that the quantum corrections, in full analogy with the above, modify the interaction potential and bring it to the form

$$V_F = G(\varphi)\bar{\psi}\psi, \tag{11.61}$$

which is a natural generalization of the expression (11.60). The function $G(\varphi)$ is a polynomial with random coefficients, by analogy with the basic potential $V(\varphi)$. Under such a choice, the fermion mass μ_F and the constant g of its interaction with the field $\phi = \varphi - \varphi_m$ turn out to depend on the number of the Universe (or, more precisely, on the position of the m-th minimum of the potential of φ_m):

$$\mu_F = G(\varphi_m); \quad g = G'_\varphi(\varphi_m)(\varphi - \varphi_m) + \cdots \tag{11.62}$$

Evidently, for any given fermion mass range $(\mu_F,\ \mu_F + \delta)$ and any function $G(\varphi)$, from the countable number of universes one can select a universe with such a potential at the minimum $V(\varphi_m)$, that the equality $\mu_F \cong G(\varphi_m)$ holds. This allows for using the same mechanism for a fine tuning of the Universe parameters, not only the vacuum energy density. Indeed, let it be necessary that the mass of a given fermion belong to the range $(\mu_F,\ \mu_F + \delta\mu_F)$. Then, in the countable number of universes with suitable vacuum energy densities we will always find a universe with a suitable value of $G(\varphi_m)$, such that the fermion mass will be inside the given range. Moreover, there is a countable number of such universes; among them we can select those suitable for life by other parameters. There occurs an automated "search" for universes admitting life by all parameters. The formation process of each of them is unique since the shape of the potential near each minimum is unique. Consequently, the formation processes of different universes should be described by different inflationary theories. A great number of realistic inflationary models are known by now. Very probably, all of them describe a certain subset of universes of our type.

Let us introduce a finite set of physical parameters ℓ_k, $k = 1, 2, \ldots, N_{\text{life}}$, necessary for creating our sort of life in the Universe. Let $\Re(\{\ell\}_n)$ denote the set of universes possessing a certain subset of n such parameters $\ell_1, \ell_2, \ldots, \ell_n$. Then the selection process for finding a universe suitable for life can be presented as follows:

$$\Re(\{\ell\}_0) \mapsto \Re(\{\ell\}_1) \mapsto \Re(\{\ell\}_2) \mapsto \cdots \mapsto \Re(\{\ell\}_{N_{\text{life}}}). \tag{11.63}$$

Each next set is a subset of the previous one but is nevertheless a countable number. If the number of parameters whose numerical values are important for the emergence of life is finite, $N_{\text{life}} < \infty$, then the process (11.63) leads us to a countable number of universes where life can emerge.

Lastly, let us discuss the Higgs mechanism of light fermion mass creation. Since the Higgs field interacts with the inflaton field (not necessarily to the first order in $\hbar$), it is obvious that the Higgs field potential should also contain an infinite series of powers of this field, a unique one for every universe. It is clear that there is a countable number of universes where the Higgs field potential has two close minima separated by a maximum of suitable height. It is also clear that the masses of fermions, emerging due to their interaction with the Higgs field in any two universes, are uncorrelated quantities. It means in turn that the quantum corrections to the initial potential (9.65) due to interaction with these fermions are also uncorrelated. Thus the postulate ($\star$) is free of contradictions.

The suggested approach also makes it possible to formulate some quantitative predictions. The main point is that the measure of the set of universes with an exact value of any of the parameters is zero. Therefore, for instance, an exact equality of the cosmological constant to zero is excluded. For the same reason there should not exist real scalar fields having potentials withe degenerate minima. The latter statement is true for any other sorts of fields in the absence of symmetries prescribed at the outset.

Appendix

A.1 A controversy between adherents of multiple universes (M) and an ultimate unified theory (U)

Where: a gravitational conference

When: coffee break

Who: two scientists

U: The whole development of science tends to unify the postulates. A success is obvious (the Standard Model, for instance). Do you want to entirely abandon all that?

M: But why do you think that this trend will always be successful? A wish to create a "theory of everything" — isn't it like a wish to make a theory of mass of the planet Earth? Well, by the way, the Standard Model has been created tens of years ago, and there is not much progress since then.

U: Well, the development of the inflationary idea is quite a significant progress. At least, we are moving in a positive direction by unifying the theories and testing their predictions experimentally. But how can you verify that there are many worlds? If, as you admit yourself, it is impossible to get to another universe, there cannot be any proof of your truth. So it does not sound better than creating a theory of the Earth's mass.

M: It should be recognized that it seems a weak point of the idea of multiple universes. But in this case, for instance, the photoelectric effect theory also has no right to exist: indeed, you cannot get into a solid body and make sure that the photon is really absorbed by the electron. In both cases one should accept indirect experimental proofs. On the other hand, your way of creating a unified theory will never lead to a final result. The next question will be: why does your Theory of Everything contain these parameter and not others? For example, if M-theory explains all the existing phenomena, the next question will be: why there are eleven or ten dimensions?

U: But who told you that the cognition process is finite? An asymptotic approximation to the Unified theory is also not the worst perspective. At least as compared with what you suggest. If all theories are possible, what is the reason to study anything?

M: Not quite so. The wealth of stars does not prevent us from studying the Sun. There are a lot of things to be understood in our own Universe. An asymptotic tendency to a unified theory is quite a worthy undertaking. But it does not promise an answer to an experimental fact: the finest selection of the unified theory parameters as though specially to provide conditions for emergence of the intelligence.

U: The unified theory must answer this question as well, although it looks trivial in the many-world concept. But still, to end with: where are all these universes situated?

M: It's really a difficult, almost philosophical question. We'll be able to answer it next to understanding where is our Universe.

A.2 Why do correct theories look elegant?

The idea of multiple universes with different properties also enables us to come to an understanding of such questions, quite far from physics, as the title of this section. Certainly, there cannot be a rigorous proof, not least because the notion of "elegance" is not strictly defined. Therefore let us begin with filling this gap. A criterion of elegance is the degree of economy of the means needed to achieve the goal: explanation of a certain set of phenomena and prediction of new ones. Let us express it by a simple formula:

$$R = N/P,$$

where R is the quality factor of a theory: the higher it is, the more elegant the theory as we see it; N is the number of explained phenomena, and P

is the amount of information (in bits) used to formulate the first principles of the theory.

Now the question can be formulated in another form: why do theories with large values of R turn out to be correct? Accepting the multiple-universe theory, we can reduce the problem of elegance to the following question: why is the fraction of universes with a large R-factor large as compared with the fraction of other universes? In this case it would be not surprising that our civilization has just emerged in such a universe.

It is reasonable to suppose that the more complex the structure of a given universe, the more rarely such universes appear (the complexity of a universe is determined by the number of bits necessary to describe it). Yes, certainly intelligent life cannot appear in quite simple universes, which make a majority. There is some complexity threshold, such that intelligence cannot emerge below it. Their number is much smaller than that of simple universes, but only these ones are of interest. In other words, if the whole set of universes is classified by complexity levels, we obtain the sequence

$$K_1,\ K_2,\ K_3, \ldots, K_U,\ K_{U+1}, \ldots$$

The index denotes the complexity level: K_1 is the number of the simplest universes (described by a smallest number of bits), K_2 is the number of those slightly more involved, etc. The number of the simplest universes where intelligence is possible is denoted by K_U; K_{U+1} is the number of more complex universes which can be inhabited by intelligence. It has been supposed that the more complex are universes, the smaller is their fraction. This means that

$$K_1 \gg K_2 \gg K_3 \gg \cdots \gg K_U \gg K_{U+1} \gg \cdots$$

Thus the probability of an intelligent civilization to find itself in a relatively simple universe is higher (and most probably much higher) than in a more complex one. And it is this circumstance that means that our Universe must be described by an elegant theory.

The assumption of the existence of extra dimensions turns out to be amazingly fruitful. On its basis, the many-worlds idea finds quite a natural explanation, as well as the fact of utmost importance, the fine tuning of the parameters of our Universe. Extra dimensions are a source of physical fields. It seems probable that a source of all phenomena and theories presented in the book is in the extra dimensions, and even such a vague fact as the elegance of correct physical theories becomes if not explained then at least more understandable. Unfortunately, the extra dimensions can be very hard to observe, for instance, due to their small size, so that their convincing experimental discovery can be a concern for the remote future.

Acknowledgements

We are thankful to our coauthors, friends and colleagues Alexander Berkov, Sergei Bolokhov, Heinz Dehnen, Júlio Cesar Fábris, Vladimir Ivashchuk, Stepan Grinyok, Sung-Won Kim, Maxim Khlopov, Rostislav Konoplich, Roman Konoplya, Jose P. S. Lemos, Vitaly Melnikov, Alexander Sakharov, Georgy Shikin, Sergey Sushkov, Igor Svadkovsky, Oleg Zaslavskii, Alexander Zhidenko, Alexei Zinger and many others for helpful discussions and support without which this book could never be written.

Bibliography

[1] L.F. Abbott, E. Fahri, and M. Wise, Phys. Lett. B **117**, 29 (1982).
[2] E. Abdalla, B. Cuadros-Melgar, A.B. Pavan, and C. Molina, Nucl. Phys. B **752**, 40–59 (2006).
[3] S.T. Abdyrakhmanov, K.A. Bronnikov and B.E. Meierovich, Grav. Cosmol. **11**, 75 (2005).
[4] F.C. Adams, Phys. Rev. D **47**, 426 (1993).
[5] S. Adler and R.B. Pearson, Phys. Rev. D **18**, 2798 (1978).
[6] I. Affleck and M. Dine, Nucl. Phys. B **249**, 361 (1985).
[7] A. Aguirre, in: *Universe or Multiverse?*, ed. B.J. Carr (Cambridge University Press, 2005).
[8] P. Aguiar and P. Crawford, Phys. Rev. D **62**, 123511 (2000).
[9] K. Akama, Lect. Notes Phys. 176, 267 (1982).
[10] K. Akama, Progr. Theor. Phys. **78**, 184 (1987).
[11] S. Alexeyev, K. Rannu, and D. Gareeva, *Possible observation sequences of Brans-Dicke wormholes*, Sov. Phys. JETP **140**, ... (2011).
[12] I. Antoniadis, Phys. Lett. B **246**, 377 (1985).
[13] O. Arias, R. Cardenas, and I. Quiros, Nucl. Phys. B **643**, 187 (2002).
[14] C. Armendáriz-Picon, Phys. Rev. D **65**, 104010 (2002).
[15] R. Arnowitt and J. Dent, Phys. Rev. D **71**, 124024 (2005).
[16] I. Arraut, D. Batic, and M. Nowakowski, arXiv: 1001.2226
[17] E. Ayon-Beato and A. Garcia, Phys. Rev. Lett. **80**, 5056 (1998).
[18] M. Banados, J. Silk, and S.M. West, Phys. Rev. Lett. **103**, 111102 (2009).
[19] T. Banks and M. O'Loughlin, Phys. Rev. D **47**, 540 (1993).
[20] N. Barbosa-Cendejas and A. Herrera-Aguilar, JHEP **0510**, 101 (2005); Phys. Rev. D **73**, 084022 (2006).
[21] C. Barceló and M. Visser, Phys. Lett. **466B**, 127 (1999).
[22] C. Barceló and M. Visser, Class. Quantum Grav. **17**, 3843 (2000); gr-qc/0003025.
[23] J.M. Bardeen, *Non-singular general-relativistic gravitational collapse.* Proc. Int. Conf. GR5, Tbilisi, 1968, p. 174.

[24] J.M. Bardeen, B. Carter, and S.W. Hawking, Commun. Math. Phys. **31**, 161 (1973).
[25] J.D. Barrow and S. Cotsakis, Phys. Lett. B **214** 515 (1988).
[26] J.D. Barrow and A.C. Ottewill, J. Phys. A **10**, 2757 (1983).
[27] D. Bazeia, F.A. Brito, and J.R. Nascimento, Phys. Rev. D **68**, 085007 (2003).
[28] R. Bean and J. Magueijo, Phys. Rev. D **66**, 063505 (2002).
[29] J.D. Bekenstein, Phys. Rev. D **7**, 949 (1973).
[30] J.D. Bekenstein, Ann. Phys. (N.Y.) **82**, 535 (1974).
[31] J.D. Bekenstein, in: *Cosmology and Gravitation*, ed. M. Novello, Atlantisciences, France, 2000, pp. 1-85.
[32] C.L. Bennett, Astrophys. J. Lett. **464** L1 (1996).
[33] W. Berej, J. Matyjasek, D. Tryniecki, and M. Woronowicz, Gen. Rel. Grav. **38**, 885 (2006).
[34] J.C. Berengut and V.V. Flambaum, *Astronomical and laboratory searches for space-time variation of fundamental constants.* arXiv: 1009.3693.
[35] O. Bergmann and R. Leipnik, Phys. Rev. **107**, 1157 (1957).
[36] P.G. Bergmann, Int. J. Theor. Phys. **1**, 25 (1968).
[37] G.D. Birkhoff and R. Langer, Relativity and Modern Physics. Harvard Univ. Press, Harvard, 1923.
[38] N.D. Birrell and P.C.W. Davies, *Quantum Fields in Curved Space.* Cambridge UP, Cambridge et al., 1982.
[39] M. Blagojevic, *Gravitation and gauge symmetries* (Institute of Physics, Bristol, 2002).
[40] U. Bleyer, M. Mohazzab, and M. Rainer, gr-qc/9508035.
[41] N.M. Bocharova, K.A. Bronnikov, and V.N. Melnikov, Vestnik MGU Fiz., Astron. No.6, 706 (1970).
[42] Christian G. Boehmer and Kevin Vandersloot, Phys. Rev. D **78**, 067501 (2008); arXiv: 0807.3042; Phys. Rev. D **76**, 104030 (2007); arXiv: 0709.2129.
[43] A. Bogojevic and D. Stojkovic, Phys. Rev. D **61**, 084011 (2000); gr-qc/9804070.
[44] N. Bostrom, *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (Routledge, New York, 2002).
[45] R.H. Boyer and R.W. Lindquist, J. Math. Phys. **8,**, 265 (1967).
[46] V.B. Braginsky and V.I. Panov, Zh. Eksp. Teor. Fiz. **61**, 873 (1971).
[47] R.H. Brandenberger, H.A. Feldman, and V.F. Mukhanov, Phys. Rep. **215**, 203 (1992).
[48] C.H. Brans, Phys. Rev. **125**, 2194 (1961).
[49] C. Brans and R.H. Dicke, Phys. Rev. **124**, 925 (1961).
[50] K.A. Bronnikov, Acta Phys. Pol. B **4**, 251 (1973).
[51] K.A. Bronnikov, Grav. Cosmol. **2**, 221 (1996); gr-qc/9703020.
[52] K.A. Bronnikov, Phys. Rev. D **63**, 044005 (2001); gr-qc/0006014.
[53] K.A. Bronnikov, Phys. Rev. D **64**, 064013 (2001); gr-qc/0104092.

[54] K.A. Bronnikov, J. Math. Phys. **43**, 6096 (2002); gr-qc/0204001.
[55] K.A. Bronnikov and M.S. Chernakova, Izv. Vuzov, Fiz. No. 9, 46 (2005); Russ. Phys. J. **48**, 940 (2005); gr-qc/0503025.
[56] K.A. Bronnikov and M.S. Chernakova, Grav. Cosmol. **13**, 1 (49), 51 (2007).
[57] K.A. Bronnikov, M.S. Chernakova, J.C. Fabris, N. Pinto-Neto, and M.E. Rodrigues, Int. J. Mod. Phys. D **17**, 25 (2008); gr-qc/0609084.
[58] K.A. Bronnikov, G. Clément, C.P. Constantindis, and J.C. Fabris, Grav. Cosmol. **4**, 128 (1998); Phys. Lett. A **243**, 121 (1998).
[59] K.A. Bronnikov, C.P. Constantinidis, R.L. Evangelista, and J.C. Fabris, Int. J. Mod. Phys. D **8**, 481 (1999).
[60] K.A. Bronnikov, H. Dehnen, and V.N. Melnikov, Phys. Rev. D **68**, 024025 (2003).
[61] K.A. Bronnikov, H. Dehnen, and V.N. Melnikov, Gen. Rel. Grav. **39**, 973 (2007).
[62] K.A. Bronnikov, A. Dobosz, and I.G. Dymnikova, Class. Quantum Grav. **20**, 3797 (2003).
[63] K.A. Bronnikov and E.V. Donskoy, Grav. Cosmol. **16**, 42 (2010); arXiv: 0910.4930.
[64] K.A. Bronnikov and E.V. Donskoy, Grav. Cosmol. **17**, 31 (2011); arXiv: 1110.6030.
[65] K.A. Bronnikov and E. Elizalde, Phys. Rev. D **81**, 044032 (2010); arXiv: 0910.3929.
[66] K.A. Bronnikov and J.C. Fabris, Phys. Rev. Lett. **96**, 251101 (2006).
[67] K.A. Bronnikov, J.C. Fabris, and A. Zhidenko, Eur. Phys. J. C **71**, 1791 (2011).
[68] K.A. Bronnikov, S.B. Fadeev, and A.V. Michtchenko, Gen. Rel. Grav. **35**, 505 (2003).
[69] K.A. Bronnikov and S.V Grinyok, in: *Inquiring the Universe. Essays to celebrate Prof. Mario Novello's jubilee* (Frontier Group, Rio de Janeiro, 2003); gr-qc/0205131.
[70] K.A. Bronnikov and S.V. Grinyok, Grav. Cosmol. **10**, 237 (2004); gr-qc/ 0411063.
[71] K.A. Bronnikov and S.V. Grinyok, Grav. Cosmol. **11**, 75 (2005); gr-qc/ 0509062.
[72] K.A. Bronnikov and A.V. Khodunov, Gen. Rel. Grav. **11**, 13 (1979).
[73] K.A. Bronnikov and Yu.N. Kireyev, Phys. Lett. A **67**, 95 (1978).
[74] K.A. Bronnikov and S.A. Kononogov, Metrologia **43**, R1 (2006).
[75] K.A. Bronnikov, S.A. Kononogov, and V.N. Melnikov, Gen. Rel. Grav. **38**, 1215 (2006).
[76] K.A. Bronnikov, S.A. Kononogov, V.N. Melnikov, and S.G. Rubin, Grav. Cosmol. **14**, 230 (2008).
[77] K.A. Bronnikov, R.V. Konoplich, and S.G. Rubin, Class. Quantum Grav. **24**, 1261 (2007).

[78] K.A. Bronnikov and M.A. Kovalchuk, in: *Problems of Gravitation Theory and Particle Theory* ed. by K.P. Staniukovich, 10th issue, Atomizdat, M., 1979; J. Phys. A: Math. Gen. **13**, 187 (1980).
[79] K.A. Bronnikov and B.E. Meierovich, Grav. Cosmol. **9**, 313 (2003).
[80] K.A. Bronnikov and V.N. Melnikov, Gen. Rel. Grav. **27**, 465 (1995); gr-qc/9406064.
[81] K.A. Bronnikov and V.N. Melnikov, Nucl. Phys. B **584**, 434 (2000); hep-th/0002200.
[82] K.A. Bronnikov and V.N. Melnikov, in: *Proceedings of the 18th Course of the School on Cosmology and Gravitation: The Gravitational Constant, Generalized Gravitational Theories and Experiments (30 April–10 May 2003, Erice).* Ed. G.T. Gillies, V.N. Melnikov and V. de Sabbata, Kluwer, Dordrecht/Boston/London, 2004, pp. 39–64; gr-qc/0310112.
[83] K.A. Bronnikov and S.G. Rubin, Phys. Rev. D **73**, 124019 (2006); Grav. Cosmol. **13**, 191 (2007).
[84] K.A. Bronnikov and G.N. Shikin, Grav. Cosmol. **8**, 107 (2002), gr-qc/0109027.
[85] K.A. Bronnikov, M.V. Skvortsova, and A.A. Starobinsky, Grav. Cosmol. **16**, 216 (2010); arXiv: 1005.3262.
[86] K.A. Bronnikov and A.A. Starobinsky, Pis'ma v ZhETF **85**, 1, 3-8 (2007); JETP Lett. **85**, 1, 1-5 (2007).
[87] K.A. Bronnikov and Sung-Won Kim, Phys. Rev. D **67**, 064027 (2003).
[88] K.A. Bronnikov and S.V. Sushkov, Class. Quantum Grav. **27**, 095022 (2010); arXiv: 1001.3511.
[89] T.A. Brun and J.B. Hartle, Phys.Rev. E **59**, 6370 (1999); quant-ph/9808024.
[90] T. Bunch and P. Davies, Proc. R. Soc. A **360**, 117 (1978).
[91] R.V. Buniy and S.D.H. Hsu, Phys. Lett. B **632**, 127 (2006); hep-th/0504003.
[92] Alexander Burinskii, *The Dirac–Kerr–Newman electron.* Grav. Cosmol. **14**, 109 (2008); hep-th/0507109.
[93] Alexander Burinskii, *Gravity versus Quantum theory: Is electron really pointlike?* arXiv: 1109.3872 (essay for GRF 2011 competition).
[94] A. Burinskii and S.R. Hildebrandt, Phys. Rev. D **65**, 104017 (2002); Grav. Cosmol. **9**, 20 (2003); Czech. J. Phys. **53** B283 (2003).
[95] Rong-Gen Cai and N. Ohta, Phys. Rev. D **74**, 064001 (2006); hep-th/0604088.
[96] P. Callin and F. Ravndall, Phys. Rev. D **70**, 104009 (2004).
[97] M. Campanelli and C.O. Lousto, Int. J. Mod. Phys. D **2**, 451 (1993).
[98] J. Campbell, *A Course of Differential Geometry* (Clarendon, Oxford, 1926).
[99] S. Carlip, gr-qc/0508072.
[100] B.J. Carr, Astrophys. J. **201**, 1 (1975).
[101] S.M. Carroll et al., Phys. Rev. D **66**, 024036 (2002); hep-th/0110149.
[102] S.M. Carroll, M. Hoffman, and M. Trodden, Phys. Rev. D **68**, 023509 (2003).

[103] E. Carugno et al., Phys. Rev. D **53**, 6863 (1996).
[104] R. Casadio, A. Fabbri, and L. Mazzacurati, Phys. Rev. D **65**, 084040 (2002).
[105] R. Casadio and L. Mazzacurati, Mod. Phys. Lett. A **18**, 651 (2003); gr-qc/0205129.
[106] C. Castro, A. Granik, and M.S. El Naschie, hep-th/0004152.
[107] A. Chamblin, H.S. Reall, H. Shinkai, and T. Shiromizu, Phys. Rev. D **63**, 064015 (2001); hep-th/0008177.
[108] K.C.K. Chan, J.H. Horne, and R.B. Mann, Nucl. Phys. B 447, 441 (1995).
[109] S. Chandrasekhar, *Mathematical Theory of black holes* (Clarendon Press, Oxford, 2006).
[110] V.M. Chechetkin, M.Yu. Khlopov, M.G. Sapozhnikov, and Ya.B. Zeldovich, Phys. Lett. B **118**, 329 (1982).
[111] T. Chiba and M. Yamaguchi, *Runaway domain wall and space-time varying* α. arXiv: 1102.0105.
[112] T. Chiba, *The constancy of the constants of Nature: updates.* arXiv: 1111.0092.
[113] J.M. Cline, S. Geon, and G.D. Moore, Phys. Rev. D **70**, 043543 (2004).
[114] A.G. Cohen and D.B. Kaplan, Phys. Lett. B **199**, 251 (1987).
[115] A.G. Cohen and D.B. Kaplan, Nucl. Phys. B **308**, 913 (1988).
[116] S. Coleman, Phys. Rev. D **15**, 2929 (1977).
[117] S. Coleman, Phys. Rev. D **16**, 1762 (1977).
[118] S. Coleman, Nucl. Phys. B **307**, 867 (1988).
[119] A.A. Coley, *Einstein Centennial Review Article*, astro-ph/0504226.
[120] E.J. Copeland, M. Sami, and S. Tsujikawa, Int. J. Mod. Phys. D **15**, 1753 (2006); hep-th/0603057.
[121] J.G. Cramer, R.L. Forward, M.S. Morris, M. Visser, G. Benford, and G.A. Landis, *Natural Wormholes as Gravitational Lenses.* Phys. Rev. D **51**, 3117 (1995).
[122] C. Csaki, J. Erlich, T.J. Hollowood, and Y. Shirman, Nucl. Phys. B **581**, 309 (2000).
[123] N. Dadhich, S. Kar, S. Mukherjee, and M. Visser, Phys. Rev. D **65**, 064004 (2002).
[124] Sergio Dain, *Geometric inequalities for axially symmetric black holes*, arXiv: 1111.3615.
[125] P. Danies, Mod. Phys. Lett. A **19**, 727 (2004).
[126] H. Dennhardt and O. Lechtenfeld, Int. J. Mod. Phys. A **13**, 741 (1998); gr-qc/9612062.
[127] N. Deruelle and M. Sasaki, Progr. Theor. Phys. **110**, 441 (2003).
[128] B.S. DeWitt, *Dynamical Theory of Groups and Fields*, Gordon and Breach, NY, 1965.
[129] B.S. DeWitt and N. Graham, eds., *The Many-Worlds Interpretation of Quantum Mechanics.* Princeton U, Princeton, 1973.
[130] Tushar Kanti Dey and Surajit Sen, *Gravitational lensing by wormholes Mod. Phys. Lett.* A **23**, 953 (2008); arXiv: 0806.4059.
[131] V.I. Dokuchaev, Phys. Usp. **34**, 447 (1991).

[132] V. Dokuchaev, Yu. Eroshenko, and S. Rubin, Grav. Cosmol. **11**, 99 (2005).
[133] V. Dokuchaev, Yu. Eroshenko, and S. Rubin, arXiv: 0709.0070.
[134] V.I. Dokuchaev and Yu.N. Eroshenko, Astron. Lett. **27**, 759 (2001); astro-ph/0202019.
[135] V.I. Dokuchaev, Yu.N. Eroshenko, and S.G. Rubin, Astron. Rep. **52**, 779–789 (2008); arXiv: 0801.0885.
[136] V.I. Dokuchaev, Yu.N. Eroshenko, S.G.Rubin, and D. A. Samarchenko, Astron. Letters **36**, 773–779 (2010); arXiv: 1010.5325.
[137] A.D. Dolgov and A.D. Linde, Phys. Lett. B **116**, 329 (1982).
[138] A. Dolgov, Phys. Rep. **222**, 309 (1992).
[139] A. Dolgov, K. Freese, R. Rangarajan, and M. Srednicki, Phys. Rev. D **56**, 6155 (1997).
[140] A. Dolgov and L. Silk, Phys. Rev. D **47**, 4244 (1993).
[141] A.D. Dolgov and S.H. Hansen, Nucl. Phys. B **54**, 408 (1999).
[142] A. D. Dolgov, F. R. Urban, Phys. Rev. D **77**, 083503 (2008); arXiv: 0801.3090.
[143] J.F. Donoghue, Phys. Rev. D **50**, 3874 (1994).
[144] A.G. Doroshkevich, N.S. Kardashev, D.I. Novikov, and I.D. Novikov, *Passage of Radiation through a Wormhole*. Astronomy Reports **52**, 616 (2008).
[145] A. Doroshkevich, J. Hansen, I. Novikov, and A. Shatskiy, *Passage of radiation through wormholes*. IJMPD **18**, 1665 (2009).
[146] B.A. Dubrovin, S.P. Novikov, and A.T. Fomenko, *Modern Geometry. Methods and Applications* (URSS, Moscow, 1988).
[147] M.J. Duff, L.B. Okun, and G. Veneziano, JHEP **03,** 023, 023 (2002), physics/0110060.
[148] J.P. Duruisseu and R. Kerner, Gen. Rel. and Grav. 15 797 (1983).
[149] B.A. Dubrovin, A.T. Fomenko, and S. P. Novikov, *Modern Geometry: Methods and Applications* (Nauka, Moscow, 1979; Springer, New York, 1985).
[150] G.R. Dvali, G. Gabadadze, and M. Porrati, Phys. Lett. B **484**, 112 (2000).
[151] I. Dymnikova, Gen. Rel. Grav. **24**, 235 (1992).
[152] I.G. Dymnikova, A. Dobozh, M.L. Filchenkov, and A. Gromov, Phys. Lett. B **506**, 351 (2001).
[153] I. Dymnikova and M. Fil'chenkov, gr-qc/0209065.
[154] D.A. Easson and R.H. Brandenberger, JHEP **0106**, 024 (2001).
[155] H. Ellis, J. Math. Phys. **14**, 104 (1973).
[156] J. Ellis, N. Kaloper, K.A. Olive, and J. Yokoyama, Phys. Rev. D **59**, 103503 (1999).
[157] H. van Elst, J.E. Lidsey, and R. Tavakol, Class. Quantum Grav. **11**, 2483 (1994).
[158] H. Everett, Rev. Mod. Phys. **29**, 454 (1957).
[159] C.W.F. Everitt et al., Phys. Rev. Lett. **106**, 221101 (2011); arXiv: 1105.3456.
[160] V. Faraoni, Class. Quantum Grav. **22**, 3235 (2005).

[161] Juan Fernandez-Gracia and Bartomeu Fiol, JHEP **0911**, 054 (2009); arXiv: 0906.2353.
[162] R.P. Feynman and A.R. Hibbs, *Quantum Mechanics and Path Integrals.* MacGraw-Hill, NY, 1965.
[163] L. Flamm, Phys. Z. **17**, 48 (1916).
[164] I.Z. Fisher, Zh. Eksp. Teor. Fiz. **18**, 636 (1948); gr-qc/9911008.
[165] H. Firouzjahi, S. Sarangi, and S.-H. Henry Tye, hep-th/0406107.
[166] B.N. Frolov, *Poincaré Gauge Theory of Gravity* (MGPU, Prometei, M., 2003).
[167] V.P. Frolov and I.D. Novikov, Phys. Rev. D **42**, 1057 (1990).
[168] V.P. Frolov and I.D. Novikov, *Black Hole Physics: Basic Concepts and New Developments,* Kluwer Acad. Publ., 1997.
[169] R. Gannouji, D. Polarski, A. Ranquet, and A.A. Starobinsky, JCAP **0609**, 016 (2006).
[170] J. Garriga and A. Vilenkin, Phys. Rev. D **64**, 023517 (2001); hep-th/0011262.
[171] C. Germani and R. Maartens, Phys. Rev. D **64**, 124010 (2001).
[172] G.W. Gibbons and S.W. Hawking, Phys. Rev. D **15**, 2738 (1977).
[173] E.B. Gliner, Dokl. AN SSSR **192**, 771 (1970).
[174] K. Gödel, Rev. Mod. Phys. **21**, 447 (1949).
[175] J.A. González, F.S. Guzmán, and O. Sarbach, arXiv: 0806.0608.
[176] M. Greene, J. Schwarz, and E. Witten, *Superstring Theory.* Cambridge UP, Cambridge, 1987.
[177] R. Gregory and R. Laflamme, Phys. Rev. Lett. **70**, 2387 (1993); hep-th/9301052; Nucl. Phys. B **428**, 399 (1994); hep-th/9494071.
[178] A.A. Grib, S.G. Mamayev, and V.M. Mostepanenko, *Vacuum Quantum Effects in Strong Fields.* Friedmann Laboratory Publ., SPb, 1994.
[179] A.A. Grib and Yu.V. Pavlov, *On the possible role of superheavy particles in the early Universe.* In: *I.Ya. Pomeranchuk and Physics at the Turn of the Century*, Proceedings of the International Conference. Eds. A.Berkov, N.Narozhny, L.Okun. World Scientific, Singapore, 2003, pp. 406–412; gr-qc/0009071.
[180] A.A. Grib and Yu.V. Pavlov, Grav. Cosmol. Suppl. **8**, 148–153 (2002); gr-qc/0206040.
[181] A.A. Grib and Yu.V. Pavlov, Int. J. Mod. Phys. A **17**, 4435 (2002); gr-qc/0211015.
[182] A.A. Grib and Yu.V. Pavlov, Grav. Cosmol. **15**, 44 (2009); arXiv: 0810.1724.
[183] A.A. Grib and Yu.V. Pavlov, Pis'ma v ZhETF **92**, 147 (2010) [JETP Lett. **92**, 125 (2010)]; Astropart. Phys. **34**, 581 (2011).
[184] A.A. Grib, Yu.V. Pavlov, and O.F. Piattella, *High energy processes in the vicinity of Kerr's black hole horizon,* arXiv: 1105.1540.
[185] Ø. Grøn and S. Hervik, *Einstein's General Theory of Relativity,* Springer, NY, 2007.
[186] C. Gundlach, R.H. Price, and J. Pullin, Phys. Rev. D **49**, 883 (1994), gr-qc/9307009.

[187] C. Gundlach, *Critical phenomena in gravitational collapse,* Living Rev. Rel. **2**, 4, 1999; gr-qc/0001046.
[188] N. Günther, D. Nicole, and D. Wallace, J. Phys. A **13**, 1755 (1980).
[189] U. Günther and A. Zhuk, *Remarks on dimensional reduction in multidimensional cosmological models*, gr-qc/0401003.
[190] U. Günther, P. Moniz and A. Zhuk, Astrophys. Space Sci. **283**, 679 (2003).
[191] A.H. Guth, Phys. Rev. D **23**, 347 (1981).
[192] A.K. Guts, *Elements of a Theory of Time* (Nasledie, Omsk, 2004, in Russian).
[193] J.C. Hafele and R.E. Keating, Science **177**, 166 and 168 (1972).
[194] M. Hamermesh, *Group theory and its application to physical problems.* Addison–Wesley, Reading, Massachusetts — Palo Alto — London, 1964.
[195] Tiberiu Harko, Zoltan Kovacs, Francisco S. N. Lobo, *Thin accretion disks in stationary axisymmetric wormhole spacetimes.* Phys. Rev. D **79**, 064001 (2009).
[196] J.B. Hartle and S.W. Hawking, Phys. Rev. D **28**, 2960 (1983).
[197] S.W. Hawking, Nature **248**, 30 (1974); Commun. Math. Phys. **43**, 199 (1975).
[198] S.W. Hawking and G.F.R. Ellis, *The Large Scale Structure of Space-Time*, Cambridge UP, Cambridge, 1973.
[199] S.W. Hawking and W. Israel, ed. General Relativity, *Einstein centenary survey.* Cambridge University Press, 2003.
[200] Sean A. Hayward, Phys. Rev. Lett. **96**, 031103 (2006); gr-qc/0506126.
[201] M. Heusler, *Black Hole Uniqueness Theorems* (Cambridge University Press, 1997).
[202] A. Hewish et al., Nature **217**, No. 5139, 709 (1968).
[203] D. Hochberg, A.A. Popov, and S.V. Sushkov, Phys. Rev. Lett. **78**, 2050 (1997).
[204] D. Hochberg and M. Visser, Phys. Rev. D **56**, 4745 (1997).
[205] D. Hochberg and M. Visser, Phys. Rev. D **58**, 044021 (1998).
[206] C.J. Hogan, in: *Universe or Multiverse*?, ed. B.J. Carr, Cambridge University Press, 2005.
[207] R. Holman, E.W. Kolb, S.I. Vadas, and Y. Wang, Phys. Rev. D **43**, 995 (1991).
[208] P. Hořava and E. Witten, Nucl. Phys. B **460**, 506 (1996); ibid. **475**, 94 (1996).
[209] Gary T. Horowitz, Phys. Scripta **T117**, 86 (2005); hep-th/0312123.
[210] S.D.H. Hsu, Phys. Lett. B **644**, 67–71 (2007); hep-th/0608175.
[211] X.M. Hu and Z.C. Wu, Phys. Lett. B **149**, 87 (1984).
[212] A. Ishibashi and H. Kodama, Prog. Theor. Phys. **110**, 901–919 (2003); hep-th/0305185.
[213] A. Ishibashi and H. Kodama, arXiv: 1103.6148, to appear in Prog. Theor. Phys. Suppl.
[214] C. Itzykson and J.B. Zuber, Quantum Field Theory. McGraw-Hill, NY, 1984.

[215] D.D. Ivanenko, P.I. Pronin, and G.A. Sardanashvili, *Gauge Theory of Gravity* (MGU Press, M., 1985, in Russian).
[216] Ted Jacobson and Thomas P. Sotiriou, J. Phys.: Conf. Ser. **222**, 012041 (2010); arXiv: 1006.1764.
[217] J.T. Jebsen, Ark. Mat. Ast. Fys. (Stockholm) 15, nr.18 (1921).
[218] Nils Voje Johansen and Finn Ravndal, Gen. Rel. Grav. **38**, 537 (2006); physics/0508163.
[219] P. Jordan, Schwertkraft und Weltall. Vieweg, Braunschweig, 1955.
[220] S. Kachru, R. Kallosh, A. Linde, and S. P. Trivedi, Phys. Rev. D **68**, 046005 (2003).
[221] N. Kaloper et. al, Phys. Rev. Lett. **85**, 928 (2000).
[222] T. Kaluza, Sitzungsber. d. Berl. Akad., 1921, S. 966.
[223] A.B. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems.* Cambridge University Press, Cambridge, 1996; Faktorial, Moscow, 1999.
[224] G.L. Kane, M.J. Perry, and A. N. Zytkow, New Astron. **7**, 45 (2002).
[225] P. Kanti, Int. J. Mod. Phys. A **19**, 4899 (2004).
[226] P. Kanti, B. Kleihaus, and J. Kunz, *Wormholes in dilatonic Einstein-Gauss-Bonnet theory.* arXiv: 1108.3003. To appear in PRL.
[227] N.S. Kardashev, I.D. Novikov, and A.A. Shatskiy, Int. J. Mod. Phys. D **16**, 909 (2007); astro-ph/0610441.
[228] R. Kerner, Gen. Rel. Grav. **14**, 453 (1982).
[229] R.P. Kerr, Phys. Rev. Lett. **11**, 237 (1963).
[230] S.V. Ketov, *Introduction to Quantum Theory of Strings and Superstrings.* Nauka, Novosibirsk, 1990, in Russian.
[231] S. Khalil and C. Muños, Contemp. Phys. **43**, 51 (2002).
[232] M.Yu. Khlopov, *Fundamentals of Cosmoparticle Physics* (URSS, M., 2004).
[233] M.Yu. Khlopov, R.V. Konoplich, R. Mignani, S.G. Rubin, and A.S. Sakharov, Astroparticle Physics **12**, 367 (2000).
[234] M.Yu. Khlopov and S.G. Rubin, *Cosmological Pattern of Microphysics in the Inflationary Universe*, vol. 144 (2004). Kluwer Academic Publishers, Dordrecht-Boston-London, series: Fundamental Theories of Physics.
[235] M. Khlopov, S. Rubin, and A. Sakharov, Phys. Rev. D **62**, 083505 (2000).
[236] N. Khviengia, Z. Khviengia, H. Lü, and C.N. Pope, Class. Quantum Grav. **15**, 759 (1998).
[237] C. Kiefer, in: *Canonical Gravity – from Classical to Quantum*, ed. by J. Ehlers and H. Friedrich, Springer, Berlin, 1994.
[238] A.A. Kirillov, Phys. Lett. B **632**, 453 (2006).
[239] A.A. Kirillov and E.P. Savelova, Phys. Lett. B **660**, 93 (2008).
[240] A.A. Kirillov and E.P. Savelova, Grav. Cosmol. **14**, 256 (2008).
[241] O. Klein, Z. Phys. **37**, 895 (1926); **46**, 188 (1927).
[242] S. Kobayashi, K. Koyama, and J. Soda, Phys. Rev. D **65**, 064014 (2002).
[243] L. Kofman, A. Linde, and A. Starobinsky, Phys. Rev. Lett. **73**, 3195 (1994).

[244] E.W. Kolb, Physica Scripta **36**, 199 (1991).
[245] E. Komatsu, Astrophys. J. Suppl. **192**, 18 (2011).
[246] S.A. Kononogov, *Metrology and Fundamental Physical Constants*. Standartinform, M., 2008, in Russian.
[247] S.A. Kononogov and V.N. Melnikov, Izmer. Tekhnika **6**, 1 (2005).
[248] N.P. Konopleva and V.N. Popov, *Gauge Fields*. Atomizdat, M., 1980, in Russian.
[249] R.V. Konoplich, Sov. Nucl. Phys. **32**, 1132 (1980).
[250] R.V. Konoplich, Teor. Mat. Fiz. **73**, 1286 (1987).
[251] R.V. Konoplich, M.Yu. Khlopov, S.G. Rubin, and A.S. Sakharov, Preprint 1203, Rome1, Rome, 1998.
[252] R.V. Konoplich and S.G. Rubin, Yad. Fiz. **37**, 1330 (1983); Yad. Fiz. **42**, 1282 (1985).
[253] R.V. Konoplich and S.G. Rubin, Yad. Fiz. **44**, 558 (1986).
[254] R.A. Konoplya and C. Molina, Phys. Rev. D **71**, 124009 (2005); gr-qc/0504139.
[255] R.A. Konoplya and A. Zhidenko, Phys. Rev. D **81**, 124036 (2010); arXiv: 1004.1284.
[256] R.A. Konoplya and A. Zhidenko, Rev. Mod. Phys. **83**, 793–836 (2011); arXiv: 1102.4014.
[257] M.Yu. Konstantinov, Int. J. Mod. Phys. D **7**, 1 (1998).
[258] S. Krasnikov, Phys. Rev. D **62**, 084028 (2000).
[259] A.B. Krebs and S.G. Rubin, Phys. Rev. B **49**, 11808 (1994).
[260] H. Kroger, G. Melkonian, and S.G. Rubin, Gen. Rel. Grav. **36**, 1649 (2004).
[261] M.D. Kruskal, Phys. Rev. **119**, 1743 (1960).
[262] L.D. Landau and E.M. Lifshitz, *Quantum Mechanics*. Izd. Fiz. Mat. Lit., M., 1963.
[263] L.D. Landau and E.M. Lifshitz, *Field Theory* (Nauka, M., 1973).
[264] G. Lazarides, hep-ph/0204294.
[265] José P.S. Lemos, F.S.N. Lobo, and S.Q. de Oliveira, Phys. Rev. D **68**, 064004 (2003).
[266] José P.S. Lemos and Oleg B. Zaslavskii Phys. Rev. D **76**, 084030 (2007); arXiv: 0707.1094.
[267] José P.S. Lemos and Oleg B. Zaslavskii Phys. Rev. D **78**, 124013 (2008); arXiv: 0811.2778.
[268] José P. S. Lemos and Vilson T. Zanchin Phys. Rev. D **83**, 124005 (2011); arXiv: 1104.4790.
[269] W. Lerche, D. Lust, and A. N. Schellekens, Nucl. Phys. B **287**, 477 (1987).
[270] A. Lichnerovicz, *Theorie globale des connexions et des groupes d'holonomie*. Edizioni Cremonese, Roma, 1955.
[271] A.D. Linde, Lett. Nuovo Cim. **39**, 401 (1984).
[272] A.D. Linde, *The Large-Scale Structure of the Universe*. Harwood Academic Publishers, London, 1990.
[273] A.D. Linde, *Particle Physics and Inflationary Cosmology*, Harvard Acad. Press, Geneva, 1990.

[274] A. Linde, Phys. Lett. B **259**, 38 (1991).
[275] A.D. Linde, Physica Scripta **36**, 35 (1991).
[276] A. Linde, Phys. Rev. D **49**, 1783 (1994).
[277] B. Liu, L. McLerran and N. Turok, Phys. Rev. D **46**, 2668 (1992).
[278] M. Livio and M.J. Rees, Science **309**, 1022 (2005).
[279] J.C. Long et al., Nature **421**, 922 (2003).
[280] V. Lukash, Sov. Phys. JETP **79**, 1601 (1980).
[281] D.H. Lyth and E.D. Stewart, Phys. Rev. D **54**, 7186 (1996).
[282] D.H. Lyth and A. Riotto, Phys. Rep. **314**, 1 (1999); hep-ph/9807278.
[283] R. Maartens, Living Rev. Rel. **7**, 7 (2004); gr-qc/0312059.
[284] K. Maeda, Phys. Rev. D **39**, 3159 (1989).
[285] K. Maeda, J.A. Stein-Schabes, and T. Futamase, Phys. Rev. D **39**, 2848 (1989).
[286] L. Magaard, PhD thesis, Kiel, 1963.
[287] G. Magnano and L.M. Sokolowski, Phys. Rev. D **50**, 5039 (1994).
[288] M. Malquarti, E.J. Copeland, A.R. Liddle, and M. Trodden, Phys. Rev. D **67**, 123503 (2003).
[289] A.S. Mazumdar and S.K. Sethi, Phys. Rev. D **46**, 5315-5320 (1992).
[290] Pawel O. Mazur, *Black Hole Uniqueness Theorems,* hep-th/0101012.
[291] Manasse R. Mbonye and Demosthenes Kazanas, Phys. Rev. D **72**, 024016 (2005); gr-qc/0506111.
[292] P.O. Mazur and E. Mottola, Proc. Nat. Acad. Sci. **101**, 9545 (2004); gr-qc/0109035.
[293] A. Melfo, N. Pantoja, and A. Skirzewski, Phys. Rev. D **67**, 105003 (2003).
[294] X.H. Meng and P. Wang, Class. Quantum Grav. **20**, 4949 (2003).
[295] M.B. Mijić, M.S. Morris, and W.M. Suen, Phys. Rev. D **39**, 1496 (1989).
[296] C.W. Misner, K.S. Thorne, and J.A. Wheeler, *Gravitation*. Freeman, San Francisco, 1973.
[297] N.V. Mitskievich, *Physical Fields in General Relativity*. Nauka, M., 1969.
[298] N.V. Mitskievich, A.P. Yefremov, and A.I. Nesterov, *Field Dynamics in General Relativity*. Energoatomizdat, M., 1985.
[299] Leonardo Modesto, arXiv: 0811.2196.
[300] Leonardo Modesto, John W. Moffat, and Piero Nicolini, Phys. Lett. B **695**, 397 (2011); arXiv: 1010.0680.
[301] Leonardo Modesto and Piero Nicolini, Phys. Rev. D **82**, 104035 (2010); arXiv: 1005.5605.
[302] M. Morris, K.S. Thorne, and U. Yurtsever, Phys. Rev. Lett. **61**, 1446 (1988).
[303] Emil Mottola, arXiv: 1107.5086.
[304] H. Mouri and Y. Taniguchi, Astroph. J. **566**, L17 (2002); astro-ph/0201102.
[305] D. Müller, H.V. Fagundes, and R. Opher, Phys. Rev. D **66**, 083507 (2002).
[306] S. Nasri, P.J. Silva, G.D. Starkman, and M. Trodden, Phys. Rev. D **66**, 045029 (2002).
[307] H.P. Nilles, Phys. Rep. **110**, 1 (1984).
[308] S. Nojiri and S.D. Odintsov, Phys. Lett. B **562** 147 (2003).

[309] S. Nojiri and S.D. Odintsov, Phys. Rev. D **68**, 123512 (2003).
[310] S. Nojiri, S.D. Odintsov, Int. J. Geom. Meth. Mod. Phys. **4**, 115 (2007); hep-th/0601213.
[311] S. Nojiri, S.D. Odintsov, and S. Ogushi, Int. J. Mod. Phys. A **17**, 4809 (2002).
[312] K. Nordtvedt, Astroph. J. **161**, 1059 (1970).
[313] I. Novikov, A.G. Polnarev, A. Starobinsky, and Ya.B. Zel'dovich, Astron. Astrophys. **80**, 104 (1979).
[314] I.D. Novikov, *The time machine and self-consistent evolutions in problems with self-interaction.* Preprint Nordita-90/38A. Nordisk Inst. f. Teoretisk Fysik, Kobenhavn, 1990.
[315] I.D. Novikov, N.S. Kardashev, and A.A. Shatskii, *The multicomponent Universe and the astrophysics of wormholes.* Physics Uspekhi **50**, 965 (2007).
[316] U. Nucamendi and M. Salgado, Phys. Rev. D **68**, 044026 (2003).
[317] H. Ochiai and K. Sato, Prog. Theor. Phys. **103**, 893 (2000).
[318] I. Olasagasti and A. Vilenkin, Phys. Rev. D **62**, 044014 (2000).
[319] K.A. Olive, M. Peloso, and J-P. Uzan, *The wall of fundamental constants.* arXiv: 1011.1504.
[320] J.R. Oppenheimer and G. Volkoff, Phys. Rev. **55**, 374 (1939).
[321] J.P. Ostriker and P.J.E. Peebles, Astroph. J. **186**, 467 (1973).
[322] M. Pavšič, Phys. Lett. A **116**, 1 (1986).
[323] A. Peltola and G. Kunstatter, Phys. Rev. D **79**, 061501 (2009); arXiv: 0811.3240.
[324] R. Penney, Phys. Rev. **182**, 1383 (1969).
[325] R. Penrose, Rev. del Nuovo Cim. **1**, 252 (1969).
[326] A.M. Perelomov, Ann. Inst. Henrí Poincaré **24**, 161 (1976).
[327] A. Polnarev and M. Khlopov, Phys. Usp. **145**, 369 (1985).
[328] V.N. Ponomarev, A.O. Barvinsky, and Yu.N. Obukhov, *Geometrodynamic Methods and Gauge Approach to the Theory of Gravitational Interactions.* Energoatomizdat, M., 1985.
[329] J. Polchinski, *String Theory,* Cambridge UP, Cambridge, 1998.
[330] R.V. Pound and G.A. Rebka, Jr., Phys. Rev. Lett. **4**, 337 (1960).
[331] R.V. Pound and J.L. Snider, Phys. Rev. **140**, B788 (1960).
[332] Alexei Pozanenko and Alexander Shatskiy, *On a possible connection between wormholes and soft gamma repeaters.* Grav. Cosmol. **16**, 259 (2010); arXiv: 1007.3620.
[333] F. Quevedo, Class. Quantum Grav. **19**, 5721 (2002).
[334] F. Rahaman, M. Kalam, and S. Chakraborty, *Gravitational lensing by stable C-field wormhole.* Chin. J. Phys. **45**, 518 (2007); arXiv: 0705.0740.
[335] R. Rajaraman, *Solitons and Instantons. An Introduction to Solitons And Instantons in Quantum Field Theory.* North-Holland, Amsterdam — New-York — Oxford, 1982.
[336] L. Randall and R. Sundrum, Phys. Rev. Lett. **83**, 3370 (1999).
[337] L. Randall and R. Sundrum, Phys. Rev. Lett. **83**, 3690 (1999).

[338] P.K. Rashevsky, *Riemannian Geometry and Tensor Analysis*. Nauka, M., 1967.

[339] M.J. Rees, *Our Cosmic Habitat*. Princeton Univ. Press, Princeton, 2002.

[340] M. Rees, R. Ruffini, and J.A. Wheeler, *Black Holes, Gravitational Waves and Cosmology: an Introduction to Current Problems*. Gordon and Breach, NY — London — Paris, 1974.

[341] M. Regis, M. Serone, and P. Ullio, Phys. Lett. B **663**, 250 (2008); hep-ph/0612286.

[342] S.-J. Rey, Nucl. Phys. B **284**, 706 (1987).

[343] A.G. Riess, Astron. J. **116**, 1009 (1998).

[344] W. Rindler, *Essential relativity*. Springer, NY, 1979.

[345] T.G. Rizzo, JHEP **0506**, 079 (2005); hep-ph/0503163, H. Davoudiasl and T. G. Rizzo, Phys. Rev. D **76**, 055009, (2007); hep-ph/0702078.

[346] V.I. Rodichev, *Gravitation Theory in an Orthogonal Hedron* (Nauka, M., 1974).

[347] E. Roessl and M. Shaposhnikov, Phys. Rev. D **66**, 084008 (2002); hep-th/0205320.

[348] Marek Rogatko, Class. Quantum Grav. **19**, L151 (2002); hep-th/0207187.

[349] Marek Rogatko, Phys. Rev. D **82**, 044017 (2010); arXiv: 1007.4374.

[350] V.A. Rubakov, Physics-Uspekhi **171**, 9, 913 (2001).

[351] V.A. Rubakov and M.E. Shaposhnikov, Phys. Lett. B **125**, 136 (1983).

[352] S. Rubin, JETP Lett. **74**, 275 (2001).

[353] S.G. Rubin, Chaos Solitons Fractals **14**, 891, (2002); astro-ph/0207013; Grav. Cosmol. **9**, 243 (2003); hep-ph/0309184.

[354] S.G. Rubin, in: *I. Ya. Pomeranchuk and Physics at the Turn of the Century*, (Moscow, 2003), p. 413; astro-ph/0511181.

[355] S.G. Rubin, Grav. Cosmol. **9**, 243 (2003).

[356] S. Rubin, M. Khlopov and A. Sakharov, Grav. Cosmol. **6**, Suppl., 51–58 (2000).

[357] S. Rubin, A. Sakharov, and M.Yu. Khlopov, Sov. Phys. JETP **92**, 921 (2001).

[358] S.G. Rubin. Sov. Phys. JETP **106**, 714 (2008).

[359] S.G. Rubin. Sov. Phys. JETP **109**, 961 (2009).

[360] J. Saavedra and Y. Vasquez, JCAP **0904**, 013 (2009); arXiv: 0803.1823.

[361] M. Safonova, D.F. Torres and G.E. Romero, *Microlensing by Natural Wormholes: Theory and Simulations*. Phys. Rev. D **65**, 023001 (2002).

[362] V. Sahni and A. Starobinsky, Int. J. Mod. Phys. D **9**, 373 (2000).

[363] V. Sahni and A. Starobinsky, Int. J. Mod. Phys. D **15**, 2105 (2006); astro-ph/0610026.

[364] Franz E. Schunck and Eckehard W. Mielke, Class. Quantum Grav. **20**, R301–R356 (2003); arXiv: 0801.0307.

[365] K. Schwarzschild, Sitzber. Deut. Akad. Wiss. Berlin, Kl. Math.-Phys. Tech., 424 (1916).

[366] S.S. Seahra and P.S. Wesson, Class. Quantum Grav. **20**, 1321 (2003); gr-qc/0302015.

[367] A. Sen, JHEP **0204**, 048 (2002); **0207**, 065 (2002).
[368] S. Shankaranarayanan and N. Dadhich, Int. J. Mod. Phys. D **13**, 1095 (2004); gr-qc/0306111.
[369] I.I. Shapiro et al., Astroph. J. **234**, L219 (1979).
[370] H.-a. Shinkai and S.A. Hayward, Phys. Rev. D **66**, 044005 (2002), gr-qc/0205041.
[371] T. Shiromizu, K. Maeda, and M. Sasaki, Phys. Rev. D **62**, 024012 (2000).
[372] Y. Shtanov, J. Traschen, and R. Brandenberger, Phys. Rev. D **51**, 5438 (1995); hep-ph/9407247.
[373] P. Sikivie, Phys. Rev. Lett. **48**, 1156 (1982).
[374] W. de Sitter, Proc. Kon. Ned. Akad Wet. **20**, 229 (1917).
[375] Anais Smailagic and Euro Spallucci, arXiv: 1003.3918.
[376] M.N. Smolyakov and I.P. Volobuev, *Linearized gravity, Newtonian limit and light deflection in RS1 model*, hep-th/0208025.
[377] L.M. Sokolowski, Acta Phys. Polon., B **39**, 2879 (2008).
[378] Euro Spallucci and Stefano Ansoldi, Phys. Lett. B **701**, 471 (2011); arXiv: 1101.2760.
[379] D.N. Spergel et al., ApJS **170**, 377 (2007); astro-ph/0603449.
[380] K.P. Staniukovich and V.N. Melnikov, *Hydrodynamics, Fields and Constants in Gravitation Theory.* Energoatomizdat, M., 1983.
[381] G.D. Starkman, D. Stojkovic, and M. Trodden, Phys. Rev. D **63**, 103511 (2001); hep-th/0012226.
[382] A. Starobinsky, JETP Lett. **30**, 682 (1979).
[383] A. Starobinsky, Phys. Lett. B **91**, 99–102 (1980).
[384] A.A. Starobinsky, Pis'ma v Astron. Zh. **7**, 67 (1981); Sov. Astron. Lett. **7**, 361 (1981).
[385] A.A. Starobinsky, Phys. Lett. B **117**,175 (1982).
[386] A.A. Starobinsky, JETP Lett. **86**, 157 (2007).
[387] A. Strominger, Phys. Rev. D **24**, 3082 (1981).
[388] W.-M. Suen and K. Young, Phys. Rev. D **39**, 2201 (1989).
[389] M. Sullivan et al., Astrophys. J. **737**, 102 (2011); arXiv: 1104.1444.
[390] Sergey V. Sushkov and Oleg B. Zaslavskii, Phys. Rev. D **79**, 067502 (2009); arXiv: 0903.1510
[391] L. Susskind, *The Anthropic Landscape of String Theory*, hep-th/0302219.
[392] J.L. Synge, Proc. Roy. Irish Acad. A **53**, 83 (1950).
[393] J.L. Synge, *Relativity: The General Theory.* North Holland, Amsterdam, 1960.
[394] P. Szekeres, Publ. Mat. Debrecen **7**, 285 (1960).
[395] F.R. Tangherlini, Nuovo Cim. **27**, 636 (1963).
[396] M. Tegmark et al., Phys. Rev. D **73**, 023505 (2006).
[397] J.M. Tejeiro and Alexis Larranaga, gr-qc/0505054.
[398] R.C. Tolman, Proc. Nat. Acad. Sci. U.S.A. **20**, 410 (1934).
[399] A.A. Tseytlin and C. Vafa, Nucl. Phys. B **372**, 443 (1992).
[400] J.-Ph. Uzan, Rev. Mod. Phys. **75**, 403 (2003).
[401] J.-Ph. Uzan, *Varying Constants, Gravitation and Cosmology. Review for Living Reviews in Relativity*, arXiv: 1009.5514.

[402] A. Vilenkin, Phys. Lett. B **117**, 25 (1982).
[403] A. Vilenkin, Phys. Rev. D **37**, 888 (1988).
[404] A. Vilenkin and L. Ford, Phys. Rev. D **26**, 1231 (1982).
[405] M. Visser, *Lorentzian wormholes.* Springer, NY–Berlin–Heidelberg, 1996.
[406] Yu.S. Vladimirov, *Reference Frames in Gravitation Theory.* Energoizdat, M., 1982.
[407] Yu.S. Vladimirov, Geometrophysics (BINOM, M., 2005).
[408] D. Vollick, Gen. Rel. Grav. **34**, 1 (2002).
[409] D. Vollick, Gen. Rel. Grav. **34**, 471 (2002).
[410] G.E. Volovik, *Proceedings of "From Quantum to Emergent Gravity: Theory and Phenomenology"* Phenomenology, Trieste, Italy, 2007 (Trieste, 2007), PoS(QG-Ph).
[411] G.E. Volovik, arXiv: 0904.1965 [gr-qc].
[412] R. Wagoner, Phys. Rev. D **1**, 3209 (1970).
[413] R. Wald, *General Relativity.* Univ. of Chicago Press, Chicago, 1984.
[414] Robert M. Wald, *Gravitational Collapse and Cosmic Censorship*, gr-qc/9710068.
[415] J.K. Webb et al., Phys. Rev. Lett. **82**, 884 (1999).
[416] J.K. Webb et al., *Evidence for spatial variation of the fine structure constant*, arXiv: 1008.3907.
[417] S. Weinberg, *Gravitation and Cosmology: Principles and Applications of GR.* Wiley, NY, 1972.
[418] S. Weinberg, Phys. Rev. Lett. **59**, 2607 (1987).
[419] S. Weinberg, *The cosmological constant problem*, astro-ph/0005265.
[420] S. Weinstein, Class. Quantum Grav. **23**, 4231 (2006).
[421] U. Weiss and W. Haeffner, Phys. Rev. D **27**, 2916 (1983).
[422] J.A. Wheeler, Phys. Rev. **97**, 511 (1955).
[423] J.A. Wheeler, in: Groups and Topology, eds. B.S. and C.M. DeWitt (Gordon and Breach, NY, 1963).
[424] J.A. Wheeler, Am. Sci. **59** 1 (1968).
[425] C. Will, *Theory and Experiment in Gravitational Physics.* Cambridge UP, Cambridge et al., 981.
[426] C. Will, Physics **4**, 43 (2011); arXiv: 1106.1198.
[427] D.L. Wiltshire, in: *Cosmology: The Physics of the Universe,* eds B. Robson, N. Visvanathan and W.S. Woolcock, World Scientific, Singapore, 1996, pp. 473–531.
[428] B. Whitt, Phys. Lett. B **145** 176 (1984).
[429] E. Witten, Nucl. Phys. B **471**, 135 (1996).
[430] P. Worden, Acta Astronautica **5**, 27 (1978).
[431] D. Yamauchi and M. Sasaki, Prog. Theor. Phys. **118**, 245 (2007); arXiv: 0705.2443.
[432] J. Yokoyama, Phys. Rev. D **58**, 083510 (1998).
[433] V. Zakharov, V. Manakov, V. Novikov, and D. Pitaevsky, *Theory of Solitons: The Inverse Problem Method.* Nauka, M., 1980, in Russian.
[434] O.B. Zaslavskii, Class. Quantum Grav. **28**, 105010 (2011).
[435] Ya.B. Zel'dovich and I.D. Novikov, Sov. Astron. **10**, 602 (1967).

[436] Ya.B. Zel'dovich and M.Yu. Khlopov, Phys. Lett. B **79**, 239 (1978).
[437] A.L. Zel'manov, *Chronomtric Invariants*. American Research Press, Rehoboth, New Mexico, 2006.
[438] A.L. Zel'manov and V.G. Agakov, *Elements of General Relativity*. Nauka, M., 1989.

Index