



Primordial Cosmology

Patrick Peter
Jean-Philippe Uzan

Primordial Cosmology

Patrick Peter
and
Jean-Philippe Uzan

Institut d'Astrophysique de Paris

Translated by
Jasna Brujić and Claudia de Rham

5410353283



OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

Edition originale: "Cosmologie primordiale"

© Editions Belin-Paris, 2005

English translation © Oxford University Press 2009

Ouvrage publié avec le concours du Ministère français
chargé de la culture – Centre national du livre

The moral rights of the authors have been asserted
Database right Oxford University Press (maker)

First Published in English 2009

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
without the prior permission in writing of Oxford University Press,
or as expressly permitted by law, or under terms agreed with the appropriate
reprographics rights organization. Enquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department,
Oxford University Press, at the address above

You must not circulate this book in any other binding or cover
and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

Printed in Great Britain
on acid-free paper by
CPI, Chippenham, Wilts

ISBN 978-0-19-920991-0

1 3 5 7 9 10 8 6 4 2

Foreword to the English edition

We live at a very special time. The universe was born about 14 billion years ago. We began to study it by very primitive tools only several thousand years ago. Less than a hundred years ago we learned about the existence of other galaxies. Finally, during the last two decades we entered the era of precision cosmology. We are able to study with incredible accuracy the structure of the observable part of the universe, finding imprints of what happened in the first milliseconds after the big bang.

But this precision would not help much without the development of new tools which previously were unavailable to cosmologists. For a long time we did not know how to study matter at densities much greater than nuclear density. This limited our ability to study physical processes in the very early universe. The standard approach was to make various assumptions about superdense matter and then study consequences of these assumptions. The situation changed in the beginning of the 1970s, when we learned that not only the temperature and density, but also the properties of elementary particles in the early universe were quite different from what we see now. In particular, according to the theory of the cosmological phase transitions, during the first 10^{-10} seconds after the big bang there was not much difference between weak and electromagnetic interactions. The next important step was the discovery of the asymptotic freedom, which implied that the strength of interactions between elementary particles decreases at large density. This allowed us to investigate physical processes very close to the big bang, at densities almost eighty orders of magnitude higher than the nuclear density.

These discoveries culminated in the invention of inflationary theory, which helped to explain why our universe is so large and flat, why it is homogeneous and isotropic, why its different parts started their expansion simultaneously. According to this theory, the early universe experienced a period of exponentially rapid expansion (inflation) in a slowly changing vacuum-like state. All elementary particles surrounding us now were produced as a result of the decay of this vacuum-like state at the end of inflation. Inflation stretched away all previously existing inhomogeneities, but simultaneously it produced new inhomogeneities from tiny quantum fluctuations amplified during the exponential growth of the universe. These fluctuations served as seeds for the subsequent process of galaxy formation.

In certain cases, these quantum fluctuations may become so large that they can be responsible not only for the formation of galaxies, but also for the formation of new exponentially large parts of the universe with different laws of low-energy physics operating in each of them. Instead of being a sphere, our universe became an eternally growing fractal, a multiverse consisting of different exponentially large parts.

In the beginning, inflationary theory could seem too radical to be true, but during the last two decades many of its predictions were confirmed by cosmological observa-

tions, and many cosmologists accepted it as a part of the new cosmological paradigm. Moreover, after the progress with vacuum stabilization in string theory achieved a few years ago, the picture of an eternally growing inflationary multiverse consisting of many exponentially large parts with different properties became a part of what is now called ‘the string theory landscape’.

It is clear that further development of cosmology will require knowledge of many subjects which traditionally did not belong to the toolbox of a cosmologist. One should know not only the general theory of relativity, but also the modern theory of elementary particles, quantum field theory, and basic elements of string theory. An additional challenge is to relate the new cosmological theory and the rapidly growing flow of observational data.

This book by Patrick Peter and Jean-Philippe Uzan successfully solves this difficult problem. It begins with an introduction to general relativity and modern particle physics, which received a powerful boost in the beginning of the 1970s. In Part II of the book they describe cosmological theory of a homogeneous and of a slightly inhomogeneous universe. Then they describe inflationary cosmology. This includes the classical theory of a homogeneous inflationary universe, the theory of inflationary perturbations, and the theory of reheating, which is responsible for the formation of all elementary particles after inflation. The authors bring it all together in Part III, where they discuss grand unification, baryogenesis, supersymmetry and supergravity, cosmological phase transitions, various models of inflation, and elements of string theory. This provides an excellent background for further independent studies of theoretical and observational cosmology and of the recent work on inflation in string theory and on the string theory landscape.

The book is well written, well illustrated and very user-friendly. I am sure that it will be appreciated not only by students who enter this exciting field of knowledge, but also by experts in cosmology who will be happy to see a unique source of information covering every important aspect of modern cosmology.

Andrei Linde
Stanford University
USA

Foreword to the French edition

Scientific cosmology was one of the most fascinating intellectual adventures of the twentieth century, and it still is in the twenty-first. It began in the early part of the century when the theoretical work of Albert Einstein, Alexandre Friedmann, and Georges Lemaître came together with the observational discoveries of Vesto Slipher, Henrietta Leavitt, and Edwin Hubble. But, for many years afterwards, the small quantity and low degree of precision of the observational data was insufficient grist to the mill for theorists, and cosmology made slow progress. This situation began to change during the latter part of the twentieth century. Some remarkable, and often unexpected, observational discoveries – like the detection of cosmic background radiation and the confirmation of the ‘dark matter problem’ – as well as the gradual increase in the precision of observational data, led to fruitful exchanges between theory and experiment. Interest in cosmology increased at the end of the century when new theoretical paradigms blossomed (e.g. cold dark matter and inflation) together with experimental discoveries that were beginning to come thick and fast, accompanied by an incredible increase in precision (e.g. anisotropies of the microwave background, the discovery of the acceleration of the rate of expansion).

At the present time, cosmology is one of the most active areas of science. In its rapid advance, it mobilizes all the resources of the latest theoretical physics (general relativity, the quantum theory of fields, superstring theory) as well as those of observation (large telescopes, satellites). The present work provides the reader with an invaluable means of access to state-of-the-art cosmology. It only requires very elementary knowledge of the basics (special relativity, non-relativistic quantum mechanics), and explains in detail all the theoretical tools underpinning modern cosmological research: general relativity, the classic and quantum theories of fields, particle physics, the theories of grand unification. It even includes an introduction to the extensions of standard theories such as, for example, supersymmetric theories, the theory of strings, and brane cosmology. Furthermore, it teaches the reader – in great detail – about the main applications of the most recent theoretical frameworks for the description of fundamental objects of cosmology: the theory of cosmological perturbations, fluctuations in the background radiation, gravitational lenses, the generation of quantum fluctuation during inflation, and topological defects.

I have no doubt that this remarkable book by Patrick Peter and Jean-Philippe Uzan will be of enormous value to a wide readership: students who are beginning research into cosmology or into astroparticle physics; experienced researchers who wish to update their knowledge; and teachers who want a synthesis of the present state of our knowledge of cosmology. All of them will find a great amount of detailed information, presented in a way that is clear and easy to understand. Finally, this is a work written by two bright young researchers that not only transmits the knowledge

that is necessary in order to understand modern theoretical cosmology and enable it to progress, but also communicates the enthusiasm generated by taking part in an exceptional intellectual adventure whose future promises to be as exciting as its past.

Thibault Damour
Institut des Hautes Études Scientifiques
Academy of Sciences
France

Acknowledgements

We would like to thank our colleagues and collaborators, in particular those from the Astrophysics Institute in Paris (Institut d’Astrophysique de Paris). The numerous discussions we shared with them have helped us in clarifying many important issues. We particularly want to thank Karim Benabed, Francis Bernardeau, François Bouchet, Tristan Brunier, Christos Charmousis, Chris Clarkson, Alain Coc, Cédric Deffayet, Nathalie Deruelle, Peter Dunsby, Ruth Durrer, George Ellis, Gilles Esposito-Farèse, Ted Jacobson, Lev Kofman, David Langlois, Jean-Pierre Lasota, Roland Lehoucq, Martin Lemoine, Keith Olive, Thiago Pereira, Nelson Pinto-Neto, Alain Riazuelo, Christophe Ringeval, Carlo Schimd, Joe Silk, Ismael Tereno, Alberto Vallinotto, Ludovic van Waerbeke, Filippo Vernizzi and Jeffrey Weeks.

This book would not have been of the same quality without the dedication of some of our colleagues in proofreading and editing of the first manuscript. We thank Nabila Aghanim, Philippe Brax, Bernard Fort, Marc Lilley, Jérôme Martin, Yannick Mellier, Renaud Parentani, Cyril Pitrou, Simon Prunet, Alain Riazuelo, Jonathan Rocher, Carlo Schimd, Frédéric Vincent, and Elisabeth Vangioni, Annick Lesne and Patrizia Castiglione for their comments and their help in putting together the final version, as well as Laurent Vigroux and the Institut d’Astrophysique de Paris for all the support in publishing this book.

We would also like to thank Andrei Linde and Thibault Damour for giving us the honour of writing the forewords to the English and French editions respectively. We also largely benefited from their useful and improving comments.

Contents

| | |
|---|-----|
| Foreword to the English edition | v |
| Foreword to the French edition | vii |
| Acknowledgements | ix |
| Introduction | 1 |
| Cosmology: an ancient yet contemporary subject | 1 |
| The specifics of cosmology | 2 |
| Primordial cosmology | 2 |
| Organization and objective of this book | 3 |
| Warnings | 5 |
| References | 6 |
| PART I OVERVIEW OF THE THEORETICAL BASIS | |
| 1 General relativity | 11 |
| 1.1 Space-time and gravity | 11 |
| 1.1.1 Absolute space and time of Newtonian physics | 11 |
| 1.1.2 Space-time of special relativity | 14 |
| 1.1.3 General relativity and curved space-time | 19 |
| 1.2 Elements of differential geometry | 21 |
| 1.2.1 Manifolds and tensors | 21 |
| 1.2.2 Geodesic equations and Christoffel symbols | 26 |
| 1.2.3 Covariant derivative, parallel transport and Lie derivative | 28 |
| 1.2.4 Curvature | 31 |
| 1.2.5 Covariant approach | 34 |
| 1.3 Equations of motion | 38 |
| 1.3.1 Einstein's equations | 38 |
| 1.3.2 Conservation equations | 41 |
| 1.3.3 ADM Hamiltonian formulation | 44 |
| 1.4 Geodesics in a curved space-time | 46 |
| 1.4.1 Conserved quantities along a geodesic | 46 |
| 1.4.2 Geodesic deviation equation | 47 |
| 1.5 Weak-field regime | 49 |
| 1.5.1 Newtonian limit | 49 |
| 1.5.2 Gravitational waves in an empty space-time | 53 |
| 1.6 Tests of general relativity | 55 |
| 1.6.1 Test of the equivalence principle | 55 |
| 1.6.2 Tests in the Solar System | 60 |

| | | |
|---|--|-----|
| 1.6.3 | Gravitational radiation | 62 |
| 1.6.4 | Need for tests at astrophysical scales | 65 |
| References | | 66 |
| 2 | Overview of particle physics and the Standard Model | 68 |
| 2.1 | From classical to quantum | 68 |
| 2.1.1 | Analytical classical mechanics | 69 |
| 2.1.2 | Quantum physics | 79 |
| 2.1.3 | Quantum and relativistic mechanics | 86 |
| 2.2 | Canonical decompositions | 89 |
| 2.2.1 | Technical clarification | 89 |
| 2.2.2 | Real free field | 90 |
| 2.2.3 | The complex scalar field | 95 |
| 2.3 | Classification and properties of the elementary particles | 98 |
| 2.4 | Internal symmetries | 101 |
| 2.4.1 | Group theory | 101 |
| 2.4.2 | Generators | 103 |
| 2.5 | Symmetry breaking | 106 |
| 2.5.1 | Gauge group | 107 |
| 2.5.2 | Higgs mechanism | 108 |
| 2.5.3 | Non-Abelian case | 110 |
| 2.6 | The standard model $SU(3)_c \times SU(2)_L \times U(1)_Y$ | 112 |
| 2.6.1 | The strong interaction (QCD) | 113 |
| 2.6.2 | Electroweak interaction | 114 |
| 2.6.3 | A complete model | 116 |
| 2.7 | Discrete invariances | 121 |
| 2.7.1 | Parity | 121 |
| 2.7.2 | Charge conjugation | 123 |
| 2.7.3 | Time reversal and CPT theorem | 124 |
| References | | 126 |
| PART II THE MODERN STANDARD COSMOLOGICAL MODEL | | |
| 3 | The homogeneous Universe | 129 |
| 3.1 | The cosmological solution of Friedmann-Lemaître | 129 |
| 3.1.1 | Constructing a Universe model | 129 |
| 3.1.2 | Cosmological and Copernican principles | 130 |
| 3.1.3 | Cosmological principle and metric | 131 |
| 3.1.4 | Kinematics | 136 |
| 3.1.5 | Space-time dynamics | 138 |
| 3.2 | Dynamics of Friedmann-Lemaître space-times | 143 |
| 3.2.1 | Some solutions | 143 |
| 3.2.2 | Dynamical evolution | 144 |
| 3.2.3 | Expansion and contraction | 148 |
| 3.3 | Time and distances | 150 |
| 3.3.1 | Age of the Universe and look-back time | 150 |

| | | |
|-------|--|-----|
| 3.3.2 | Comoving radial distance | 151 |
| 3.3.3 | Angular distances | 152 |
| 3.3.4 | Luminosity distance | 154 |
| 3.3.5 | Volume and number counts | 157 |
| 3.4 | Behaviour at small redshifts | 158 |
| 3.4.1 | Deceleration parameter | 158 |
| 3.4.2 | Expression of the time and distances | 158 |
| 3.5 | Horizons | 160 |
| 3.5.1 | Event horizon | 160 |
| 3.5.2 | Particle horizon | 162 |
| 3.5.3 | Global properties | 163 |
| 3.6 | Beyond the cosmological principle | 167 |
| 3.6.1 | Classifying space-times | 168 |
| 3.6.2 | Universe with non-homogeneous spatial sections | 170 |
| 3.6.3 | Universe with homogeneous spatial sections | 171 |
| | References | 175 |
| 4 | The standard Big-Bang model | 177 |
| 4.1 | The Hubble diagram and the age of the Universe | 177 |
| 4.1.1 | The Hubble constant | 178 |
| 4.1.2 | The contribution of supernovæ | 182 |
| 4.1.3 | The age of the Universe | 186 |
| 4.2 | Thermodynamics in an expanding Universe | 188 |
| 4.2.1 | Equilibrium thermodynamics | 188 |
| 4.2.2 | Out-of-equilibrium thermodynamics | 197 |
| 4.2.3 | Two limiting cases | 201 |
| 4.3 | Primordial nucleosynthesis | 204 |
| 4.3.1 | Main stages of the mechanism | 205 |
| 4.3.2 | Initial state | 206 |
| 4.3.3 | Freeze-out of the weak interaction and neutron to proton ratio | 208 |
| 4.3.4 | Abundances of the light elements | 211 |
| 4.3.5 | Observational status | 215 |
| 4.4 | The cosmic microwave background radiation | 216 |
| 4.4.1 | Recombination | 217 |
| 4.4.2 | Properties of the cosmic microwave background | 221 |
| 4.4.3 | Another proof of the expansion of the Universe | 226 |
| 4.5 | Status of the Big-Bang model | 226 |
| 4.5.1 | A good standard model... | 226 |
| 4.5.2 | ... but an incomplete model | 228 |
| 4.5.3 | Conclusion | 234 |
| | References | 235 |

Contents

| | |
|---|-----|
| The inhomogeneous Universe | 238 |
| 5.1 Newtonian perturbations | 238 |
| 5.1.1 Case of a static space | 238 |
| 5.1.2 Case of an expanding space | 240 |
| 5.1.3 Predictions and observables | 246 |
| 5.1.4 Towards the non-linear regime | 249 |
| 5.2 Gauge invariant cosmological perturbation theory | 251 |
| 5.2.1 Perturbed space-time | 251 |
| 5.2.2 Description of matter | 255 |
| 5.2.3 Choosing a gauge | 257 |
| 5.2.4 Einstein equations: derivation | 260 |
| 5.2.5 Einstein equations: SVT decomposition | 262 |
| 5.2.6 Perturbed conservation equation for a fluid | 264 |
| 5.2.7 Interpretation of the perturbation equations | 265 |
| 5.3 Evolution | 266 |
| 5.3.1 Vector and tensor modes | 267 |
| 5.3.2 Evolution of the gravitational potential | 269 |
| 5.3.3 Scalar modes in the adiabatic regime | 272 |
| 5.3.4 Mixture of several fluids | 278 |
| 5.4 Power spectrum of density fluctuations | 282 |
| 5.4.1 Two equivalent approaches | 283 |
| 5.4.2 Different regimes | 285 |
| 5.4.3 Some refinements | 292 |
| 5.5 The large-scale structure of the Universe | 296 |
| 5.5.1 Observing the large-scale structure | 296 |
| 5.5.2 Structures in the non-linear regime | 299 |
| References | 304 |
| The cosmic microwave background | 307 |
| 6.1 Origin of the cosmic microwave background anisotropies | 307 |
| 6.1.1 Sachs-Wolfe formula | 307 |
| 6.1.2 Angular power spectrum | 311 |
| 6.2 Properties of the angular power spectrum | 316 |
| 6.2.1 Large angular scales | 317 |
| 6.2.2 Intermediate scales | 318 |
| 6.2.3 Small scales | 323 |
| 6.2.4 Other effects | 326 |
| 6.3 Kinetic description | 329 |
| 6.3.1 Perturbed Boltzmann equation | 330 |
| 6.3.2 Gauge invariant expressions | 339 |
| 6.3.3 Thomson scattering and polarization | 342 |
| 6.3.4 Numerical integration | 355 |
| 6.4 Anisotropies of the cosmic microwave background | 356 |
| 6.5 Effects of the parameters on the angular power spectrum | 364 |
| 6.5.1 Cosmological parameters | 364 |
| 6.5.2 Parameters describing the primordial physics | 367 |

| | |
|---|-----|
| References | 371 |
| 7 Gravitational lensing and dark matter | 374 |
| 7.1 Gravitational lensing and its applications | 374 |
| 7.1.1 Gravitational lensing in the thin-lens regime | 375 |
| 7.1.2 Lensing by galaxies and galaxy clusters | 387 |
| 7.1.3 Gravitational distortion by the large-scale structure | 395 |
| 7.1.4 Cosmic convergence | 399 |
| 7.1.5 Cosmic shear | 402 |
| 7.1.6 Measurement of the cosmic shear | 405 |
| 7.1.7 Lensing on the cosmic microwave background | 408 |
| 7.2 Evidence for the existence of dark matter | 411 |
| 7.2.1 Dark matter in galaxies | 411 |
| 7.2.2 Dark matter in clusters and groups of galaxies | 419 |
| 7.2.3 Cosmological evidence | 422 |
| 7.2.4 Summary | 423 |
| 7.2.5 Candidates and constraints | 423 |
| References | 445 |
| 8 Inflation | 450 |
| 8.1 Genesis of the paradigm | 451 |
| 8.1.1 Original motivations | 451 |
| 8.1.2 Resolution of the Big-Bang problems | 452 |
| 8.1.3 First models of inflation | 454 |
| 8.1.4 Inflation as a de Sitter phase | 455 |
| 8.2 Dynamics of single-field inflation | 457 |
| 8.2.1 Equations of evolution | 457 |
| 8.2.2 Slow-roll parameters | 459 |
| 8.2.3 End of inflation | 463 |
| 8.2.4 Some examples | 465 |
| 8.2.5 Classification of inflationary models | 469 |
| 8.3 Quantum fluctuations during inflation | 470 |
| 8.3.1 Massless test scalar field in a de Sitter space-time | 470 |
| 8.3.2 Massive test field in de Sitter | 476 |
| 8.3.3 Massive test field during slow-roll inflation | 477 |
| 8.4 Quantum fluctuations of the inflaton | 478 |
| 8.4.1 Overview of the questions to address and expected results | 479 |
| 8.4.2 Perturbed quantities | 479 |
| 8.4.3 Perturbation equations | 481 |
| 8.4.4 Evolution of the long-wavelength modes | 483 |
| 8.4.5 Junction conditions and their applications | 485 |
| 8.4.6 Quantization of the density perturbations | 488 |
| 8.4.7 Gravitational waves | 492 |
| 8.5 Perturbations in the slow-roll regime | 494 |
| 8.5.1 An exact solution: power-law inflation | 495 |
| 8.5.2 General case | 496 |

| | | |
|--|--|-----|
| 8.5.3 | Relation to observations | 501 |
| 8.5.4 | Reconstruction of the potential | 505 |
| 8.6 | End of inflation and reheating | 507 |
| 8.6.1 | Perturbative reheating | 507 |
| 8.6.2 | Theory of preheating | 510 |
| 8.7 | Eternal inflation | 516 |
| 8.7.1 | Heuristic argument | 516 |
| 8.7.2 | Stochastic approach | 518 |
| 8.8 | Extensions | 523 |
| 8.8.1 | Multifield inflation | 523 |
| 8.8.2 | Non-Gaussianity | 530 |
| 8.8.3 | Trans-Planckian problem | 532 |
| 8.9 | Status of the paradigm | 537 |
| | References | 541 |
| PART III BEYOND THE STANDARD MODELS | | |
| 9 | Grand unification and baryogenesis | 549 |
| 9.1 | Interactions | 549 |
| 9.1.1 | Superposition principle | 550 |
| 9.1.2 | N -particle states | 551 |
| 9.2 | Grand unification | 554 |
| 9.2.1 | Problems of the standard model | 554 |
| 9.2.2 | Unification of the coupling constants | 560 |
| 9.2.3 | Unification models | 562 |
| 9.2.4 | Consequences of grand unification | 566 |
| 9.2.5 | Neutrino mass | 568 |
| 9.3 | Baryogenesis | 569 |
| 9.3.1 | Sakharov conditions for baryogenesis | 569 |
| 9.3.2 | Electroweak anomalies | 571 |
| 9.3.3 | Electroweak baryogenesis | 573 |
| 9.3.4 | Leptogenesis | 575 |
| 9.3.5 | Affleck–Dine mechanism | 576 |
| | References | 578 |
| 10 | Extensions of the theoretical framework | 579 |
| 10.1 | Scalar-tensor theory of gravity | 579 |
| 10.1.1 | Formulation | 580 |
| 10.1.2 | Local constraints | 584 |
| 10.1.3 | Cosmological aspects | 584 |
| 10.1.4 | Phenomenological aspects | 588 |
| 10.1.5 | $f(R)$ gravity and scalar-tensor theory | 593 |
| 10.2 | Quantum field theory in curved space-time | 594 |
| 10.2.1 | Quantum physics, classical gravity | 594 |
| 10.2.2 | Particle creation | 597 |
| 10.2.3 | A complete example | 602 |

| | |
|---|-----|
| 10.3 Supersymmetry and supergravity | 604 |
| 10.3.1 Technical generalities | 604 |
| 10.3.2 Wess–Zumino model | 610 |
| 10.3.3 Gauge field | 614 |
| 10.3.4 Supersymmetry breaking | 615 |
| 10.3.5 The minimal supersymmetric standard model (MSSM) | 618 |
| 10.3.6 Supergravity | 621 |
| References | 625 |
| 11 Phase transitions and topological defects | 628 |
| 11.1 Phase transitions | 628 |
| 11.1.1 Thermal field theory | 628 |
| 11.1.2 Dynamics of the symmetry breaking | 630 |
| 11.1.3 Formation of topological defects | 633 |
| 11.2 Domain walls | 635 |
| 11.2.1 Correlation length | 635 |
| 11.2.2 Static configurations | 638 |
| 11.3 Vortices | 640 |
| 11.3.1 Kibble mechanism for cosmic strings | 641 |
| 11.3.2 Internal structure | 643 |
| 11.4 Monopoles | 647 |
| 11.5 Textures | 648 |
| 11.6 Defects in general | 648 |
| 11.6.1 Connectedness | 649 |
| 11.6.2 Fundamental group | 649 |
| 11.6.3 Homotopy group | 650 |
| 11.6.4 Semi-topological defects | 651 |
| 11.7 Walls in cosmology | 653 |
| 11.7.1 Distribution and evolution | 653 |
| 11.7.2 Observational constraints | 655 |
| 11.8 Cosmological monopoles | 656 |
| 11.8.1 GUT monopoles are unavoidable | 656 |
| 11.8.2 The monopole problem | 656 |
| 11.8.3 Possible solutions to the monopole problem | 660 |
| 11.9 Cosmic strings | 661 |
| 11.9.1 General properties | 661 |
| 11.9.2 Gravitational effects of strings | 662 |
| 11.9.3 Effect of a network on the microwave background | 666 |
| 11.9.4 Other consequences of cosmic strings | 671 |
| References | 674 |
| 12 Cosmological extensions | 678 |
| 12.1 Construction of inflationary models | 678 |
| 12.1.1 A consistent model: supergravity | 678 |
| 12.1.2 <i>F</i> -term inflation | 681 |
| 12.1.3 <i>D</i> -term inflation | 683 |

| | | |
|--------|---|-----|
| 12.2 | Cosmological constant and dark energy | 684 |
| 12.2.1 | The cosmological constant problem | 685 |
| 12.2.2 | The nature of dark energy | 687 |
| 12.2.3 | Quintessence | 690 |
| 12.2.4 | Other models | 697 |
| 12.2.5 | Other approaches | 701 |
| 12.2.6 | Parameterization of the equation of state | 705 |
| 12.2.7 | Implications for the formation of the large-scale structure | 708 |
| 12.3 | Varying constants | 711 |
| 12.4 | Topology of the Universe | 716 |
| 12.4.1 | Local and global structures | 716 |
| 12.4.2 | Mathematical introduction | 718 |
| 12.4.3 | Classification of the three-dimensional manifolds | 718 |
| 12.4.4 | Observational signatures | 726 |
| | References | 733 |
| 13 | Advanced topics | 737 |
| 13.1 | Extra dimensions: Kaluza–Klein theory | 737 |
| 13.1.1 | General relativity in D dimensions | 737 |
| 13.1.2 | Projected equations in four dimensions | 739 |
| 13.1.3 | Dilaton and Einstein–Maxwell theory | 741 |
| 13.1.4 | Einstein frame | 742 |
| 13.1.5 | Compactification | 743 |
| 13.2 | A few words on string theory | 744 |
| 13.2.1 | From particles to strings | 744 |
| 13.2.2 | Superstrings | 746 |
| 13.2.3 | Open and closed strings | 749 |
| 13.2.4 | Dualities | 752 |
| 13.2.5 | Low-energy Lagrangians | 753 |
| 13.2.6 | Origin of three space-like dimensions | 755 |
| 13.3 | The Universe as a ‘brane’ | 758 |
| 13.3.1 | Motivations | 758 |
| 13.3.2 | Induced Einstein equations | 761 |
| 13.3.3 | The Randall–Sundrum model | 764 |
| 13.3.4 | Cosmological phenomenology | 767 |
| 13.3.5 | Possible extensions | 768 |
| 13.3.6 | The Universe as a defect | 769 |
| 13.3.7 | Models of induced gravity | 771 |
| 13.4 | Initial singularity and a bouncing Universe | 772 |
| 13.4.1 | Approaching the singularity | 772 |
| 13.4.2 | The pre-Big-Bang scenario | 775 |
| 13.4.3 | The cyclic scenarios | 781 |
| 13.4.4 | Regular bounce and power spectrum | 785 |
| | References | 793 |

| | |
|---|-----|
| Appendix A Numerical values | 796 |
| A.1 Physical constants | 796 |
| A.2 Astrophysical quantities | 796 |
| A.3 Units | 797 |
| A.3.1 Natural units | 797 |
| A.3.2 Conversion factors | 798 |
| A.4 Particle physics | 798 |
| A.5 Cosmological quantities | 799 |
| A.6 Electromagnetic spectrum | 799 |
| References | 803 |
| Appendix B Special functions | 804 |
| B.1 Euler functions | 804 |
| B.2 Spherical harmonics | 805 |
| B.2.1 Definition | 805 |
| B.2.2 Expressions | 805 |
| B.2.3 Properties | 805 |
| B.2.4 Fourier transform | 806 |
| B.2.5 Useful integrals | 807 |
| B.3 Bessel functions | 808 |
| B.3.1 Definition | 808 |
| B.3.2 Asymptotic properties | 808 |
| B.3.3 Special cases | 809 |
| B.3.4 Spherical Bessel functions | 809 |
| B.3.5 Some useful integrals | 810 |
| B.4 Legendre polynomials | 811 |
| B.4.1 Associated Legendre polynomials | 811 |
| B.4.2 Legendre polynomials | 811 |
| B.5 Fourier transform and the eigenmodes of the Laplacian | 812 |
| B.5.1 Fourier transform | 812 |
| B.5.2 Power spectrum | 812 |
| B.5.3 Cartesian coordinates | 813 |
| B.5.4 Spherical coordinates | 813 |
| References | 815 |
| Appendix C Useful cosmological quantities | 816 |
| C.1 Background space | 816 |
| C.1.1 Geometry | 816 |
| C.1.2 Matter | 817 |
| C.2 Perturbed quantities | 819 |
| C.2.1 Geometry | 819 |
| C.2.2 Matter | 821 |
| Index | 823 |

Introduction

Cosmology: an ancient yet contemporary subject

Over the past few decades, the study of the Universe has undergone great transformations, both theoretical and observational. Cosmology, whose aim is to understand the global properties and large-scale structures of the Universe (their origin, evolution, characteristics ...) has now entered an era of precision.

For many years, cosmology had been a very speculative subject, relying as much on metaphysics as on physics. The development in 1915 of the theory of general relativity gave birth to a consistent theoretical framework, making it possible to mathematically formulate the notion of space and time. It was then possible, as early as in 1924, to formulate cosmological models grounded within this theory. Such models of the Universe, whose main characteristic is to be expanding, enabled many observational features such as the recession of the galaxies, as pointed out by Edwin Hubble, to be understood. For about twenty years, cosmology confined itself to the description and the reconstruction of this expansion, as well as to debating whether this expansion was real.

Within a second period, starting around 1948, studying physical processes in an expanding space led to the formulation of the hot Big Bang model. Taking this expansion seriously, one can deduce from the laws of nuclear physics that the Universe has a thermal history. In particular, the light nuclei must have been formed within the first minutes of the expansion, and there must exist an electromagnetic background radiation. Thanks to these developments, the model stands on three solid foundations, including the confirmation of the expansion. These conclusions only rely on the hypothesis that general relativity is a valid description of the dynamics of the Universe and on the validity of ordinary physics. No other alternative model has to date been able to reproduce these observations, especially not with so few ingredients.

Since the 1980s, many developments have brought the Big Bang model to a new level. Not only does it describe the dynamics of the Universe on large scales, but it also attacks questions on the properties and the origin of the large-scale structures. Among the most recent progress, one can mention an explosion of observational data:

- since the discovery of the cosmic microwave background in 1965, its observation has become more and more precise, from the initial discovery of the thermal anisotropies by the satellite COBE (COsmic Background Explorer) (1992) to the full-sky map by the satellite WMAP (Wilkinson Microwave Anisotropy Probe) in 2003;
- the Hubble diagram, relating the recession speed of galaxies to their redshift, could be extended to large distances, mainly through the observation of distant supernovæ;

2 Introduction

- the new galaxy catalogues now have several thousands of objects, which makes it possible to map the three-dimensional distribution of matter;
- since 2000, new observations have appeared, such as the cosmic shear.

One can notice several major theoretical developments:

- the formulation of the paradigm of inflation that links the origin of the large-scale structure of the Universe to high-energy physics;
- the formulation of the cold dark-matter model that represents a framework for the study of the large-scale structure;
- a considerable number of indications and observational evidence leading to the postulate of the existence of dark matter and dark energy. The study of their nature will shape research within the coming decades.

The specifics of cosmology

Despite these developments and the abundance of observations, cosmology nonetheless maintains a different status compared to the other sciences. Only one Universe is indeed observable, and, moreover, from a unique position in space and time. The main part of these observations are therefore confined to our past light cone. Cosmology can therefore not prevent the use of hypotheses that are impossible to check, such as the *cosmological principle*, which has strong implications concerning cosmological space-time symmetries. The elaboration of a cosmological model relies mainly on three hypotheses: (1) the choice of a theory of gravity, (2) hypotheses concerning the nature of the matter present in the Universe and (3) a symmetry hypothesis.

One should stress that cosmological observations cannot be read independently of these theoretical hypotheses (Figs. 1 and 2). Therefore the validity of a model cannot be proven, but one can look for a consistency between observations and the theoretical framework within which they are read. This does not prevent cosmology from excluding models that give predictions that are incompatible with observations, as in any other areas of physics.

Primordial cosmology

Primordial cosmology, which will be discussed in this book, relies on two pillars: numerous astronomical observations made by more and more capable instruments that reveal the local structure of our Universe with an ever-increasing precision, and a set of high-energy physics theories that are expected to describe the dynamics of the primordial Universe.

We therefore have to deal with a two-way process (Fig. 2). On the one hand (arrow labeled '1'), using some high-energy physics hypotheses, one can build a phenomenological model for the primordial Universe. This model must provide the properties of the primordial fluctuations that give birth, through gravitational instability, to the large-scale structure of the Universe. We should therefore be able to predict the resulting properties of the Universe. Comparing a specific model with different sets of observations allows us to constrain it. On the other hand (arrow labeled '2'), these models bring solutions to the problems encountered while interpreting the observations: the nature of dark matter, of dark energy, etc.

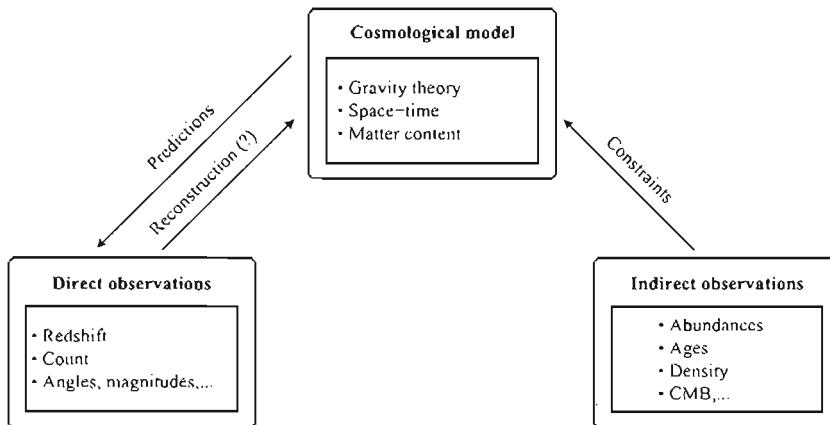


Fig. 1 The construction of a cosmological model and its link with observations. Most observations cannot be interpreted outside this model.

Thanks to these theories, we hope to be able to reproduce the observations, and on the other hand, to use these observations to constrain the theories when they are extrapolated to experimentally inaccessible energies.

Organization and objective of this book

The aim of this book is to provide the tools for primordial cosmology, to describe its state and the current questions. It is organized into three parts. The first part reviews the useful basis for the study of cosmology. The second part focuses on the standard cosmological model. It represents the heart of this book and we have tried to describe in detail all the technical aspects. The third part describes the extensions of this standard model and sheds light on the subjects currently under discussion.

The first part reviews the required theoretical notions. Indeed, since standard cosmology is based on the theories of general relativity and on the standard model of particle physics, Chapters 1 and 2 will provide an overview of these two areas. We try to distinguish between what is well established and the murkier areas of the current theories. The different extensions of these standard frameworks play an important role for the elaboration of models of the primordial Universe. It is thus important to demarcate the limits of these areas.

The second part presents the standard cosmological model.

Chapters 3 and 4 draw up the most common cosmological solutions of general relativity and the hot Big Bang model. Chapter 3 is mainly mathematical while Chapter 4 focuses on the physics. These two parts consider only a homogeneous and isotropic Universe without taking account of the structure present in the Universe. This standard model will be our reference model. We will discuss its successes and its problems.

The study of the large-scale structure of the Universe is developed in Chapter 5, which is the undeniable technical heart of primordial cosmology. Cosmological perturbation theory describes how initial perturbations evolve in an expanding Universe.

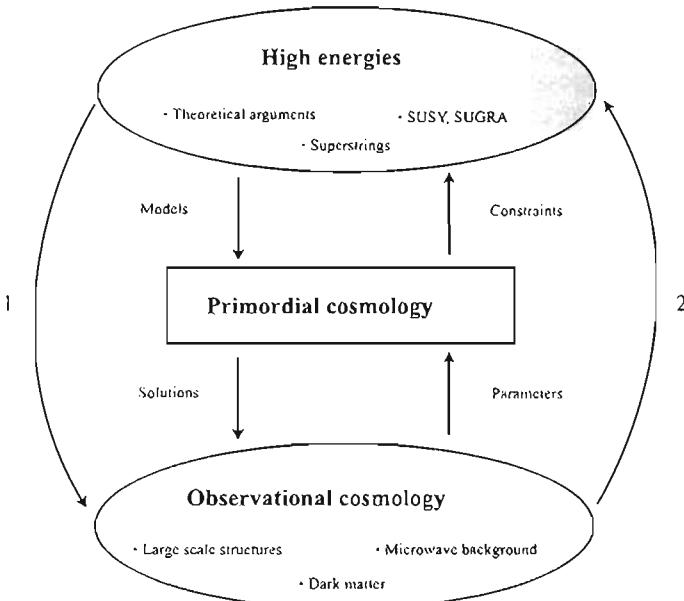


Fig. 2 Primordial cosmology constructs phenomenological models that are inspired from high-energy physics in order to describe the Universe. It looks for mechanisms able to describe the primordial phases of our Universe, the origin of large-scale structures and of the matter content of our Universe. Each model must produce observational predictions. In return, observational cosmology makes it possible to measure the parameters of these models and to set, in the best cases, some constraints on the initial theories.

We will establish the equations of this theory and discuss its solutions within some regimes.

Our book does not focus much on the observational aspects of cosmology, nevertheless Chapters 6 and 7 are dedicated to the computation of two fundamental observational predictions. Our aim is to make the reader able to compute these predictions within a given model. Chapter 6 describes the computation of the angular power spectrum of the cosmological microwave background temperature anisotropies, and Chapter 7 focuses on gravitational lensing effects and on the proofs of the existence of dark matter.

Chapter 8 develops the inflation paradigm. This paradigm is fundamental to explain the origin of the large-scale structure and it appears to be not only consistent with all observations, but also to be the only mechanism so far capable of reproducing these observations. The hot Big Bang model associated with inflation provides our reference model of the Universe. This chapter describes the general properties of inflationary models and analyses in detail the computation of the primordial fluctuations spectrum, mainly within the slow-roll approximation. We will also address the reheating mechanisms that make the transition between the inflationary Universe, the hot Big Bang, and the idea of eternal inflation.

The third part addresses different extensions of these standard formalisms, for both the fundamental theories and the cosmological framework itself.

Chapter 9 focuses on grand unified theories and on their consequences for the understanding of the origin of ordinary matter, the so-called baryogenesis.

Chapter 10 describes different extensions of the theoretical framework that have consequences for cosmology. We first describe the scalar-tensor theories of gravity. These theories represent the simplest extension of general relativity and include, besides the graviton, a scalar interaction. These models are motivated by numerous high-energy theories, such as string theory, for which the low-energy action is usually a scalar-tensor theory. We then describe the basis of quantum field theory in curved space-time. This will place the computations made in Chapter 8 on more solid ground. Finally, we will describe supersymmetry and supergravity as extensions to the standard model of particle physics. These theories have many candidates for dark-matter particles and have many scalar fields, which makes them relevant for the elaboration of inflationary models.

Chapter 11 addresses the issue of phase transitions during the primordial Universe and describes models of topological defects generated during phase transitions with symmetry breaking. These relics can play an important role during the early Universe even if we now know that they cannot be responsible for the origin of the large-scale structure as we observe it.

The two last chapters describe the different implications of these extensions for cosmology. Chapter 12 draws the consequences from supersymmetry for the construction of inflationary models. We will then address the question of the origin of the acceleration of our Universe. Chapter 13 focuses on the phenomenology of string theory for the primordial Universe. These subjects are still currently debated by researchers and our aim is to sketch their motivations and characteristics.

Warnings

In order to avoid an overflow of references in the text, we have made the choice of citing in this book mainly detailed reviews and texts that we judge to be clear and pedagogical or where a complete bibliography can be found.

As for the numerical values of the cosmological parameters deduced from different analyses, these values should be taken with great care, as they depend: (1) on the set of chosen observations, (2) on the model we compare with these observations and for instance on the number of parameters allowed to vary, and (3) on which kind of statistical analysis has been used. The values we quote should be considered more as rough indications rather than the latest values to remember. We also point out that observations are developing rapidly such that some conclusions, for instance on dark matter, may change on the time scale of a year.

References

For a discussion on the background of the formulation of cosmological models:

- [1] G.F.R. ELLIS, 'Cosmology and verifiability', *Q. Jl. Astr. Soc.* **16**, 245–264, 1975.
- [2] H. BONDI, *Cosmology*, Cambridge Monograph on Physics, 1961.
- [3] G.F.R. ELLIS, 'Issues in the philosophy of cosmology', [astro-ph/0602280](#).

Introductory books for cosmology:

- [4] E.R. HARRISON, *Cosmology, the science of the Universe*, Cambridge University Press, 2000.
- [5] J. BERNSTEIN, *Introduction to cosmology*, Englewood Cliffs, 1995.
- [6] A. LIDDLE, *An introduction to modern cosmology*, John Wiley and Sons, 1999.
- [7] B. RIDDEN, *Introduction to cosmology*, Addison Wesley, 2003.
- [8] S. WEINBERG, *The first three minutes*, Bantam Books, 1977.
- [9] E.R. HARRISON, *Darkness at night: a riddle of the Universe*, Harvard University Press, 1987.

Books of the history of modern cosmology:

- [10] H. KRAUGH, *Cosmology and controversy*, Princeton University Press, 1996.
- [11] J.P. LUMINET, Alexandre Friedmann, Georges Lemaître, *Essais de cosmologie*, Seuil, 1997.

Reference books in cosmology:

- [12] P.J.E. PEEBLES, *Physical cosmology*, Princeton University Press, 1971.
- [13] P.J.E. PEEBLES, *Principle of physical cosmology*, Princeton University Press, 1993.
- [14] S. WEINBERG, *Gravitation and cosmology: principles and applications of the general theory of relativity*, John Wiley and Sons, 1972.
- [15] E.W. KOLB and M.S. TURNER, *The early Universe*, Addison Wesley, 1993.
- [16] T. PADMANABHAN, *Structure formation in the Universe*, Cambridge University Press, 1993.
- [17] S. DODELSON, *Modern cosmology*, Academic Press, 2003.
- [18] J. RICH, *Fundamentals of cosmology*, Springer, 2001.
- [19] J.A. PEACOCK, *Cosmological physics*, Cambridge University Press, 1998.
- [20] A.R. LIDDLE and D.H. LYTH, *Cosmological inflation and large-scale structure*, Cambridge University Press, 2000.
- [21] V.F. MUKHANOV, *Physical foundations of cosmology*, Cambridge University Press, 2005.

For astrophysical complements:

- [22] T. PADMANABHAN, *Theoretical astrophysics* (3 volumes), Cambridge University Press, 2002.
- [23] J. BINNEY and S. TREMAINE, *Galactic dynamics*, Princeton University Press, 1987.
- [24] M.S. LONGAIR, *Galaxy formation*, Springer-Verlag, 1998.

Part I

Overview of the theoretical basis

1

General relativity

Since gravity is the only long-range force that cannot be screened, the properties of the Universe on large scales will be essentially determined by this force.

The goal of this chapter is not to present a complete and rigorous description of the theory of general relativity. Such detailed explanations exist in the literature and we refer the reader to them for more details; see Refs. [1–5]. Here we want to recall the basic hypotheses and formulas of general relativity that are necessary for the rest of this book.

We recall in Section 1.1 the evolution of the mathematical concept of space-time from Newtonian physics to general relativity. Section 1.2 summarizes the main definitions of the mathematical tools used in general relativity. We derive Einstein equations that determine the dynamics of space-time and the conservation equations that determine the evolution of matter in Section 1.3 and we show in Section 1.5 how Newtonian gravity is recovered in a weak-field limit. Section 1.4 describes the kinetics in curved space-time and we summarize the various tests of general relativity in Section 1.6. The other fundamental interactions, such as electromagnetic, weak and strong interactions, are discussed in chapter 2.

1.1 Space-time and gravity

1.1.1 Absolute space and time of Newtonian physics

In Newtonian physics, space is described by an absolute and immutable mathematical space. This space is Euclidean in 3 dimensions. We can therefore endow it with an origin and with 3 arbitrary reference axes, thus determining an absolute frame of reference.

Time is also ideal and absolute. It is independent of the motion of any observer and plays the role of an external parameter. For any event P , there exists an intuitive notion of simultaneity (defined as a set of all events that occur at the same time). If we consider a second event Q there are three possible outcomes: (1) one can in principle go from P to Q , Q therefore belongs to the future of P , (2) one can in principle go from Q to P , Q therefore belongs to the past of P and (3) it is impossible to travel between P and Q , the two events are therefore simultaneous. This causal structure of Newtonian space-time is represented in Fig. 1.1.

Since space is assumed Euclidean, Pythagoras' theorem would permit us to calculate the distance between two neighbouring points

$$d\ell^2 = (dx^1)^2 + (dx^2)^2 + (dx^3)^2, \quad (1.1)$$

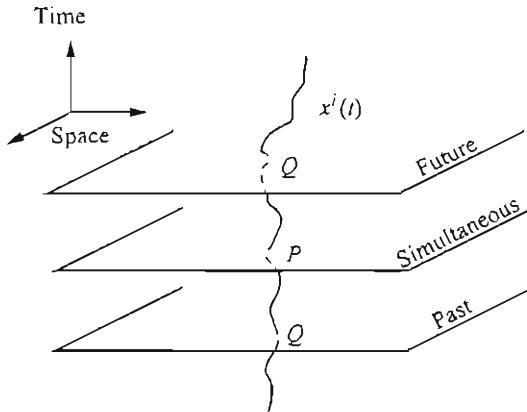


Fig. 1.1 The causal structure of Newtonian space-time. If Q can be related to P by a trajectory $x^i(t)$ with finite speed ($dx^i/dt < \infty$), Q is in the past or the future of P , otherwise it is simultaneous. The set of all simultaneous events is a three-dimensional Euclidean space.

in Cartesian coordinates. We can rewrite this distance in shorter form as

$$d\ell^2 = \sum_{i=1}^3 (dx^i)^2 = \sum_{ij} \delta_{ij} dx^i dx^j \equiv \delta_{ij} dx^i dx^j, \quad (1.2)$$

where δ_{ij} is the Kronecker symbol equal to 1 if $i = j$ and 0 otherwise, and the latin indices $i, j \dots = 1 \dots 3$. This form introduces the Einstein summation convention according to which we implicitly assume a sum over all repeated indices. For example, if T and V are two vectors, $T \cdot V \equiv T_i V^i = \delta_{ij} T^i V^j = T^1 V^1 + T^2 V^2 + T^3 V^3$ is the scalar product of those two vectors.

The trajectory of any body is therefore given in the parametric form by $x^i(t)$. The travel time from point A at t_A to point B at t_B is given by $t_A - t_B$ and is independent of the trajectory between A and B .

Laws of physics do not necessarily require Cartesian coordinates. One can, for example, use spherical or cylindrical coordinates if they are better adapted to the problem at hand. Let us suppose that we have a system of coordinates (y^i) related to Cartesian coordinates (x^j) by a relationship of the form $x^j(y^i)$. The vector dx of coordinates dx^i in the Cartesian system has coordinates $dy^i = (\partial y^i / \partial x^j) dx^j$; we therefore deduce that the distance between two neighbouring points, in any coordinate system, is

$$d\ell^2 = g_{ij}(y^k) dy^i dy^j, \quad g_{ij}(y^k) = \frac{\partial x^m}{\partial y^i} \frac{\partial x^n}{\partial y^j} \delta_{mn}, \quad (1.3)$$

g_{ij} being the metric of space in the new coordinates. For example, spherical coordinates (r, θ, φ) are defined by

$$x = r \sin \theta \cos \varphi, \quad y = r \sin \theta \sin \varphi, \quad z = r \cos \theta.$$

Applying (1.3), we deduce that the only non-zero components of the metric in spherical coordinates are

$$g_{rr} = 1, \quad g_{\theta\theta} = r^2, \quad g_{\varphi\varphi} = r^2 \sin^2 \theta.$$

The line-element (1.2) takes the form

$$ds^2 = dr^2 + r^2 (d\theta^2 + \sin^2 \theta d\varphi^2).$$

1.1.1.1 Galilean group

The choice of the reference frame is arbitrary. It is thus interesting to exhibit the frame transformations that preserve the form (1.1) of the line element $d\ell^2$. It is equivalent to finding the coordinate transformations keeping the 3 axes of reference orthogonal. The set of these transformations is therefore given by the rigid rotations of these three axes and a translation of the origin

$$x^i \rightarrow y^i = \Lambda_j^i(t)x^j + T^i(t), \quad \Lambda_j^i(t)\Lambda_k^j(t) = \delta_k^i. \quad (1.4)$$

The rotation Λ_j^i can be time dependent, but it is rigid because it is independent of the position x^j ; it is therefore the same for all points at a given time. T^j is a time-dependent translation. The invariance under the action of these transformations (1.4) reflects the isotropy and the homogeneity of Euclidean space. Among all these rigid transformations, those of the Galilean group play a central role. The Galilean group is the subgroup of time-independent rotations and of translations linear in time, $T^i = T_0^i - v^i t$, where T_0^i and v^i are constant vectors, hence preserving the equivalence between inertial frames. The laws of physics are assumed to be identical for all inertial reference frames, which means that they are invariant under the action of a transformation of the Galilean group

$$x^i \rightarrow y^i = \Lambda_j^i x^j + T_0^i - v^i t, \quad \Lambda_j^i \Lambda_k^j = \delta_k^i. \quad (1.5)$$

This group is composed of three rotations, three translations and three changes of the inertial reference frames. Taking also the possibility to change the origin of time, it is therefore ten parameters group (three angles, a vector with three translation coordinates, one velocity vector and a change in the origin of time).

1.1.1.2 Trajectory of a massive body

The equations of motion of a body of mass m evolving in a gravitational potential ϕ can be obtained by determining the extremum of the Lagrangian constructed from the total energy of the particle

$$L = \frac{1}{2} m_i v^2 + m_G \phi = \left[\frac{1}{2} m_i g_{ij}(x^k) \dot{x}^i \dot{x}^j + m_G \phi(x^k) \right], \quad (1.6)$$

where we have considered that the *inertial mass*, m_i , and the *gravitational mass*, m_G , could a priori be different. The Euler-Lagrange equations $[\delta L / \delta x^i] = d(\delta L / \delta \dot{x}^i) / dt$ are rewritten in any coordinate system as¹

¹See the following chapter for a brief reminder of analytical mechanics and Ref. [5] for a complete and detailed study.

$$\ddot{x}^i + \Gamma_{jk}^i \dot{x}^j \dot{x}^k = -\frac{m_c}{m_i} g^{ij} \partial_j \phi, \quad \Gamma_{jk}^i = \frac{1}{2} g^{ip} (\partial_j g_{kp} + \partial_k g_{jp} - \partial_p g_{jk}). \quad (1.7)$$

In Cartesian coordinates, $g_{ij} = \delta_{ij}$ and $\Gamma_{ij}^k = 0$. We therefore recover the equations of motion in their classical form $\ddot{x} = -\nabla \phi$, when $m_i = m_c$. Symbols Γ_{ij}^k represent the fictitious forces (centrifugal, Coriolis...) that have to be taken into account when the reference frame is not inertial.

Also, note that (1.7) is independent of the mass of the body if $m_i = m_c$; all bodies fall identically, whatever their mass or chemical composition. This is what is called the *weak equivalence principle* or *universality of free fall*. Actually, in the Newtonian case, it is a priori possible to introduce two different masses and the equality between the gravitational mass and the inertial mass seems fortuitous. Using pendulums, Newton (1686) showed that

$$\frac{|m_c - m_i|}{m_c + m_i} < 10^{-3}. \quad (1.8)$$

The equality between the gravitational and the inertial mass was therefore only required up to one thousandth. However, Newton considered this equality to be the crux of his theory of gravitation and he dedicated the introduction of his *Principia* to it.

As we will see, this property plays a central role in the construction of theories of gravitation.

1.1.2 Space-time of special relativity

The Maxwell laws of electromagnetism have the property of not being invariant under the Galilean group (1.5). For example, the electric force generated by a charge at rest is not invariant since a magnetic force appears in a reference frame where the source is in inertial motion (see, for example, Ref. [6]). On the other hand, Maxwell deduced from his theory that light is an electromagnetic wave whose velocity of propagation, c , depends on the permittivity and the permeability of the medium. One had to introduce a fictitious medium in which these waves could propagate; the aether. The Maxwell equations would only be valid in this particular reference frame, identified with the absolute reference system of Newton. In a different inertial reference frame the speed of light should therefore have been measured as $c \pm v$, which opened up the possibility of determining the absolute reference frame, that of the aether, at the price of abandoning the principle of Galilean relativity.

1.1.2.1 Lorentz transformations

The experiments by Michelson and Morley (1887) refuted the above-mentioned hypothesis; light was shown to have the same speed of propagation in any given inertial reference frame. In 1905, Einstein changed this perspective by reaffirming the principle of *special relativity* according to which all the laws of nature should have the same form, no matter what the inertial reference frame is (see for example Ref. [7] for the historical aspects). In particular, this should have been the case for the Maxwell equations, which implied that the speed of light should be the same in all inertial reference

frames. Consequently, one deduced that the rules for changing the inertial frame can not be those of Galileo (1.5).

The new group of transformations is the Lorentz group. An inertial observer can label an event by considering a rigid Euclidean reference frame, which allows him to determine the Cartesian coordinates x, y, z . At every point in that grid, he can place a synchronized clock in such a way that all events are labelled by a quadruplet (t, x, y, z) . A second observer in an inertial reference frame moving at a velocity v along the Ox axis with respect to the first observer can also construct such a coordinate system. It permits him to identify the same events by another quadruplet (t', x', y', z') . The two sets of coordinates are related by the Lorentz transformation, called a *boost*,

$$\begin{aligned} ct' &= \frac{ct - (v/c)x}{\sqrt{1 - v^2/c^2}}, \\ x' &= \frac{x - (v/c)ct}{\sqrt{1 - v^2/c^2}}, \\ y' &= y, \\ z' &= z. \end{aligned} \quad (1.9)$$

These transformations have the property of reducing to Galilean transformations when $v \ll c$ and keep the speed of light equal to c in all inertial reference frames.

1.1.2.2 Minkowski space-time

Special relativity postulates that all inertial observers are equivalent, the sets of coordinates determined by different observers have no intrinsic meaning. The invariant quantity upon changing the reference frame is obtained by considering a pseudo-Euclidean version of Pythagoras' theorem. The line-element (1.1) is replaced by the line-element between two events, in Minkowski coordinates, of space-time

$$ds^2 = -(dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2 \equiv \eta_{\mu\nu} dx^\mu dx^\nu, \quad (1.10)$$

with $x^0 \equiv ct$. The Greek indices vary between 0 and 3 and the Einstein summation convention is extended to these values. We note that contrary to the Euclidean metric, ds^2 can be negative or vanish for non-coincident events. In this framework, space and time are united into a space-time, the *Minkowski space-time*. This space-time is absolute, just as in the Newtonian framework. It is endowed with four orthonormal axes constituting the absolute Minkowski reference frame.

The set of points such that $ds^2 = 0$ is known as the light cone and represents, for all points M in the space-time, the ensemble of points that can receive a light beam emitted in M (future light cone) or received in M (past light cone). This light cone determines the causal structure of Minkowski space-time illustrated in Fig. 1.2. Indeed, since the velocity of all bodies is smaller than c , only pairs of points (M, P) such that $ds^2 < 0$, in other words such that M is at the interior of the light cone in P , can be joined by trajectories emanating from P . The distance is then time-like and the proper time τ is generally introduced by the relation

$$ds^2 = -c^2 d\tau^2. \quad (1.11)$$

16 General relativity

If $ds^2 > 0$, the interval is said to be space-like and the points of the couple (M, P) cannot be joined by any physical trajectory.

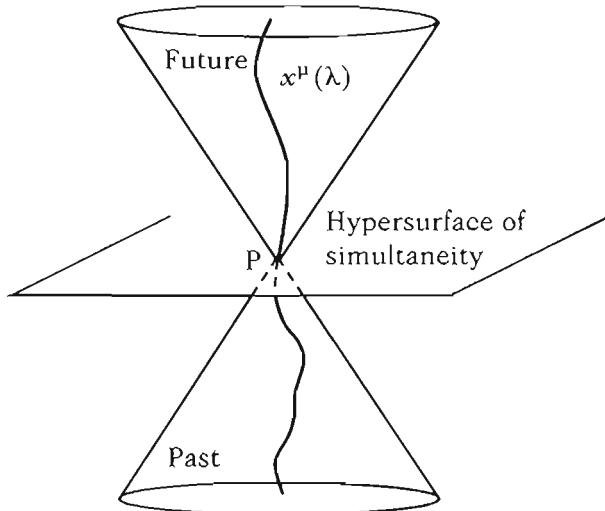


Fig. 1.2 The causal structure of Minkowski space-time. The notion of simultaneity is replaced by the notion of a light cone defined by $ds^2 = 0$. The worldline $x^\mu(\lambda)$ emanating from P is found at the interior of the light cone ($ds^2 < 0$). All points exterior to the cone can never be joined by a physical trajectory leaving from P . They are separated from P by a space-like interval ($ds^2 > 0$).

1.1.2.3 Proper time

The trajectory of any massive particles can be represented in the parametric form $x^\mu(\lambda)$, where λ is an arbitrary parameter (increasing towards infinity). We can reparameterize this worldline as a function of the proper time

$$\tau = \int \sqrt{-\eta_{\mu\nu} U^\mu U^\nu} d\lambda,$$

where $U^\mu = dx^\mu/d\lambda$ is the tangent vector to the worldline. In this new parameterization, the tangent vector is given by $u^\mu = dx^\mu/d\tau$ and it satisfies

$$u^\mu u_\mu = -c^2.$$

Let us now consider an observer O' that moves with a velocity v with respect to an inertial reference frame, S , where an inertial observer O is located. As O is motionless, its trajectory is given by $x^i = \text{const}$ and its proper time is

$$d\tau_O^2 = dt^2.$$

The moving observer O' has a trajectory such that $dx^i/dt = v^i$ with respect to the reference frame S . Its proper time is thus

$$d\tau_{O'}^2 = dt^2 - \delta_{ij} dx^i dx^j / c^2 = \left(1 - \frac{v^2}{c^2}\right) dt^2.$$

In conclusion, the proper times measured by two observers O and O' , obtained by eliminating the time coordinate, are related by

$$d\tau_{O'} = \sqrt{1 - \frac{v^2}{c^2}} d\tau_O. \quad (1.12)$$

Thus, contrary to the Newtonian case, the duration of the journey measured by O is not the same as that measured by O' , depending on their trajectories. If the clocks of O and O' have been initially synchronized, they will not be synchronous at another meeting point. This is what is called the phenomenon of *time dilation*.

1.1.2.4 Poincaré group

Just as in the Newtonian case, the metric (1.10) can be written in any coordinate system (y^μ) that is related to Minkowski coordinates (x^ν) by a relation of the form $x^\nu(y^\mu)$. A vector dx^ν in Minkowski space has the coordinates $dy^\mu = (\partial y^\mu / \partial x^\nu) dx^\nu$. We thus deduce that the line-element (1.10) has the form

$$ds^2 = g_{\mu\nu}(y^\lambda) dy^\mu dy^\nu, \quad g_{\mu\nu}(y^\lambda) = \frac{\partial x^\alpha}{\partial y^\mu} \frac{\partial x^\beta}{\partial y^\nu} \eta_{\alpha\beta}. \quad (1.13)$$

For example, we can consider the Rindler coordinates, related to Minkowski coordinates by

$$t = R \sinh T, \quad x = R \cosh T, \quad y = y', \quad z = z'.$$

The line-element (1.13) then takes the form

$$ds^2 = -R^2 dT^2 + dR^2 + dy'^2 + dz'^2.$$

The laws of physics should be, according to the principle of special relativity, independent of any preferred inertial reference frame. It is therefore important to exhibit transformations of the reference frame that preserve the form (1.10). These particular transformations constitute the Poincaré group and take the form

$$x^\mu \rightarrow y^\mu = \Lambda_\nu^\mu x^\nu + T^\mu, \quad \Lambda_\nu^\mu \Lambda_\sigma^\nu = \delta_\sigma^\mu, \quad (1.14)$$

where the rotation matrices Λ_ν^μ and the vector T^μ do not depend on the coordinates. Note that this group of transformations is more constrained than the Galilean group in which the rotations could depend on time. This Poincaré group consists of 3 rotations, three Lorentz boosts and four translations. It therefore has 10 degrees of freedom, which means that 10 values are needed to completely describe any transformation of the group.

1.1.2.5 Poincaré algebra

The Poincaré group can be constructed using infinitesimal transformations. Let us first consider a simple transformation of vector $T^\mu = \delta x^\mu$ and write the transformation (1.14) for a scalar function $f(x)$ in the form $f(y^\mu) = (1 - i\delta x^\alpha P_\alpha) f(x^\mu)$. This relation defining the generator P_μ , we get $P_\mu = i\partial_\mu$ by identification with a first-order Taylor expansion of the function $f(y)$. Equivalently, a Lorentz transformation can be written $y^\mu = x^\mu + \omega_\nu^\mu x^\nu$, with $\omega_{\nu\mu} = -\omega_{\mu\nu}$, and the generator $J_{\mu\nu}$ is obtained using the definition $f(y^\mu) = (1 - i\omega^{\alpha\beta} J_{\alpha\beta}) f(x^\mu)$, which gives $J_{\mu\nu} = i(x_\mu \partial_\nu - x_\nu \partial_\mu)$.

Other representations of the Poincaré group are possible, but all of them respect the following commutation relations, which form the Poincaré algebra:

$$\boxed{[P_\alpha, P_\beta] = 0, \quad [J_{\mu\nu}, P_\alpha] = i(g_{\mu\alpha} P_\nu - g_{\nu\alpha} P_\mu),} \quad (1.15)$$

for the translations, and

$$\boxed{[J_{\mu\nu}, J_{\alpha\beta}] = i(g_{\mu\alpha} J_{\beta\nu} + g_{\mu\beta} J_{\alpha\nu} + g_{\nu\alpha} J_{\mu\beta} + g_{\nu\beta} J_{\alpha\mu}),} \quad (1.16)$$

which is the Lorentz algebra, itself independently closed. Let us recall that the commutator between two operators is defined by

$$[A, B] = AB - BA.$$

1.1.2.6 Trajectory of a free body

In an analogous way to the Newtonian case, the trajectory of a massive free body can be obtained by extremising the action

$$L = \frac{1}{2} g_{\mu\nu}(x^\lambda) \dot{x}^\mu \dot{x}^\nu, \quad (1.17)$$

where a dot above a variable represents a derivative with respect to λ . Euler–Lagrange equations then take the form

$$\dot{u}^\mu + \Gamma_{\nu\lambda}^\mu u^\nu u^\lambda = 0, \quad \Gamma_{\nu\lambda}^\mu = \frac{1}{2} g^{\mu\sigma} (\partial_\nu g_{\lambda\sigma} + \partial_\lambda g_{\nu\sigma} - \partial_\sigma g_{\nu\lambda}), \quad (1.18)$$

where $u^\mu = dx^\mu/d\lambda$ and $\dot{u}^\mu = du^\mu/d\lambda$. In Minkowski coordinates, we find that an inertial observer must follow a trajectory of constant velocity

$$\dot{u}^\mu = 0 \iff u^\nu \partial_\nu u^\mu = 0. \quad (1.19)$$

Note the analogy with the massive body case in (1.6), but also the difference: in the case of (1.17), it is not possible to simply introduce a term equivalent to the gravitational potential.

1.1.3 General relativity and curved space-time

Special relativity reconciles the theory of electromagnetism and the principle of relativity at the price of replacing the Galilean principle of relativity by the principle of special relativity (because it is only applicable to the inertial reference frames).

However, another contradiction occurs. The theory of Newtonian gravity introduces instantaneous action at a distance, which is incompatible with the causal structure of special relativity.

1.1.3.1 Equivalence principle

Einstein based his analysis on the fact that in Newtonian gravity all test objects fall in exactly the same way in an external gravitational field, whatever their mass or chemical composition.

In Galilean and Newtonian physics, this *universality of free fall* comes from the equality between the inertial mass and the gravitational mass, the *equivalence principle* [cf. (1.8)]. This equality may seem accidental, but it is verified experimentally with high accuracy. In the Newtonian case, a small deviation from this equality would not be catastrophic. Is this equivalence principle a good approximation, a simple coincidence or does it signal something more profound?

Einstein considered the study of accelerated motion. At first glance, it could seem that objects are following different laws of physics, since one needs to add fictitious forces (inertia, Coriolis...) to describe their dynamics. For Einstein, these forces are just as real as the force of gravity since the latter can also be eliminated in a free-fall reference frame. Thus, an accelerated observer is subject to the same laws of physics as those encountered in the gravitational field, introduced in addition to the other forces. Everything happens as if one could free oneself from gravity by an astute choice of the reference frame.

This is possible only if the equivalence principle is strictly valid. Einstein therefore elevated it to the status of a first principle. This property teaches us that the ‘confusion’ between gravity and acceleration is not accidental: it is a fundamental property of gravity. Note that this property only applies to gravity and to no other force of nature (in the case of electromagnetism, for instance, the acceleration in a given electric field would depend on the ratio between charge and mass, which varies from particle to particle).

The equivalence principle ensures that one can always find locally a reference frame in which the force of gravity is eliminated.

1.1.3.2 Tidal forces and space-time curvature

It seems as though gravity has disappeared. In fact, as we emphasized, such an operation is possible only locally. As an example, let us consider a body A orbiting around the Earth at a distance r_A . Its Newtonian acceleration is $\mathbf{a}_A = -G_N M_{\oplus} \mathbf{r}_A / r_A^3$. A neighbouring body B , orbiting according to $\mathbf{r}_B = \mathbf{r}_A + \delta \mathbf{r}$ experiences a relative acceleration with respect to the body A

$$\delta \mathbf{a}_{AB} = -\frac{G_N M_{\oplus}}{r_A^3} \left(\delta \mathbf{r} - 3 \frac{\mathbf{r}_A \cdot \delta \mathbf{r}}{r_A^2} \mathbf{r}_A \right)$$

20 General relativity

at first order in δr . This relative acceleration cannot be cancelled out. If $r_A \cdot \delta r = 0$ body B approaches body A and if $r_A \cdot \delta r = \pm r_A \delta \tau$, it moves away. Thus, a small sphere in orbit would slowly deform into a type of ellipsoidal cigar, while conserving its volume. This distortion characterizes the tidal forces.

We have seen that in the Newtonian and Minkowskian frameworks, the trajectory of a free particle corresponds to a path that minimizes the length of a trajectory between two points. Even though we can eliminate gravity locally, neighbouring trajectories have a tendency to converge or diverge from each other. In a flat space-time, only one body can free itself from gravity and the fictitious forces. However, in curved space-time it is possible to consider that all the bodies are free and follow the lines of minimum length. Gravity and fictitious forces disappear for all bodies at the expense of postulating a new property of space.

In this framework, all bodies follow the lines of shortest path and gravity has locally vanished for each one of them. However, since this is possible only locally, neighbouring geodesics reveal the curvature of space-time. Gravity has therefore not disappeared, it is simply hidden in the curvature of the space-time, which explains the tidal forces in a novel way.

This heuristic example shows us that by introducing a curved space-time we can eliminate all the fictitious forces related to the choice of a non-inertial reference frame, at the expense of introducing a curvature for the space. The trajectory of all free-falling bodies, i.e. those subject to gravity alone, can be obtained as a geodesic, a line of shortest path, in curved space-time. Accordingly, all reference frames, inertial or not, are placed on the same footing, which allows us to reconcile gravity and special relativity. This construction is however only possible because acceleration does not depend on the mass of the particles in free fall, which provides a glimpse into the relationship between the universality of free fall and the geometrization of gravity.

1.1.3.3 Gravity as a manifestation of geometry

Einstein's equivalence principle is at the heart of all metric theories of gravitation, which includes amongst others the theory of general relativity. It is based on three conditions:

- the *weak equivalence principle* (or the universality of free fall) according to which the trajectory of a neutral test body is independent of its internal structure and its composition. This body must have a negligible binding gravitational energy and be sufficiently small such that the inhomogeneities of the gravitational field can be ignored;
- the *local position invariance* according to which the result of all non-gravitational experiments is independent of the point in space-time where the experiment took place;
- the *local Lorentz invariance* according to which the results of non-gravitational experiments are independent of the motion of the laboratory as long as it is free falling.

One can argue (see, for example, Ref. [8]) that if the Einstein equivalence principle is valid then gravitation is the physical manifestation of a curved space-time, i.e. a

metric theory. Such a theory has the following three properties:

- the geometry of space-time is described by a metric,
- free bodies follow the geodesics of that geometry,
- in a local reference frame in free fall, the laws of physics take the same form as in special relativity.

This implies that the contribution of the binding energies of the three non-gravitational interactions to the mass is the same for a gravitational mass and an inertial mass.

It could be asked whether this property can be generalized to the gravitational binding energy itself. This leads to the formulation of the *strong equivalence principle*. If the gravitational binding energy contributes equally to the inertial and gravitational mass then it is possible to eliminate the external gravitational field. This can be achieved in a locally inertial reference frame in which all the laws of Nature, gravitation included, have the same form as in the absence of an external field. This principle seems poorly defined and coarse: it implies a gravitational law that has as yet not been defined. A theory of gravitation that satisfies this principle is necessarily non-linear since the gravitational field ‘weights’ and generates a secondary gravitational field, etc. General relativity and the Nordström scalar theory are examples of theories that satisfy this principle (see Ref. [8]). Since these principles are central to the construction of general relativity, it is indeed important to test them (see Section 1.6).

1.2 Elements of differential geometry

The heuristic argument of the preceding paragraph shows that we can describe the motion of a free-falling body (that is, a body on which no other force but gravity is acting) by a geodesic in a four-dimensional curved space-time. Space-time is therefore described by a four-dimensional continuum so that one would need four values to locate each event, just as in special relativity.

In the Newtonian framework and in special relativity, this is assumed to be globally valid so that an ensemble of events can be mapped bijectively with \mathbb{R}^4 . In general relativity we do not want to make such a strong hypothesis about the structure of space-time before having determined it. This situation is analogous to the description of a sphere in two dimensions: we need two numbers to characterize the position of all points. Locally, the sphere looks like a plane \mathbb{R}^2 but globally it does not.

The space-time of general relativity will therefore be modelled as a four-dimensional manifold, that is, a space with four dimensions that locally looks like \mathbb{R}^4 but not necessarily globally (like in the case of a sphere in two dimensions).

1.2.1 Manifolds and tensors

Here, we define the notions of manifolds and tensors in a formal way. The two following sections can eventually be skipped and the reader can directly go to the rules of tensorial calculus derived in the third section.

1.2.1.1 Manifolds

To define a manifold, let us recall that an open set of \mathbb{R}^n is a set that can be defined as the finite union of open balls of \mathbb{R}^n . Such a ball of radius r centred at $y = (y_1, \dots, y_n)$

is defined as the set of points x such that $|x - y| < r$.

A manifold is a space consisting of neighbourhoods that are locally like \mathbb{R}^n and that can be continuously glued together. More precisely, a manifold \mathcal{M} is a collection $\{O_i\}$ of sets that satisfy the following three properties:

1. for all points p of \mathcal{M} , there exists at least one set O_i containing p , which means that $\{O_i\}$ covers \mathcal{M} ,
2. for all i , there exists a homeomorphism ψ_i from O_i into a subset U_i of \mathbb{R}^n , that is a continuous map that associates all points p of O_i to n -tuple of real numbers that we shall call coordinates of p in the map ψ_i ,
3. if two sets O_i and O_j have a non-empty intersection, then the map $\psi_j \circ \psi_i^{-1}$ that maps the points of $\psi_i(O_i \cap O_j) \subset U_i$ into $\psi_j(O_i \cap O_j) \subset U_j$ is an infinitely differentiable function on \mathbb{R}^n .

Examples of manifolds are the Euclidean plane \mathbb{R}^2 , which needs only one chart (identity), or a sphere that requires two charts.

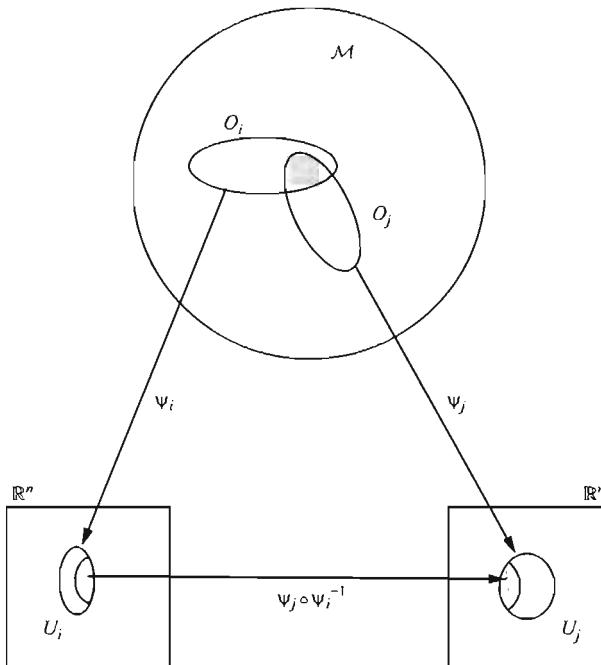


Fig. 1.3 Illustration of the relationship between different charts of the same manifold. Locally, one can relate \mathcal{M} and \mathbb{R}^n but globally this is only possible if we use a collection of charts (an atlas) that cover all the manifold.

1.2.1.2 Vectors

The Euclidean space in Newtonian physics is a vector space. However, this structure is lost in curved spaces, such as that of a sphere, but can be recovered in the limit of

infinitesimal displacements.

In \mathbb{R}^n there is a bijective map between vectors and directional derivatives and each vector $y = (y^1, \dots, y^n)$ defines a derivative $y^\mu(\partial/\partial x^\mu)$, in any arbitrary coordinate system (x^μ) . In the manifold M , consider the set F of C^∞ functions of M in \mathbb{R} . We can then define the tangent vector y at a point p in M as a map of F into \mathbb{R} , which is linear and satisfies the Leibniz rule

1. $y(af + bg) = ay(f) + by(g)$ for all f, g in F and all a, b in \mathbb{R} ,
2. $y(fg) = f(p)y(g) + g(p)y(f)$.

The set of tangent vectors at point p forms a vector space V_p [since $(y_1 + y_2)(f) = y_1(f) + y_2(f)$ and $(ay)(f) = ay(f)$] of the same dimension as the manifold. We can construct a basis $\{X_\mu\}$ of V_p by considering the map from F to \mathbb{R} defined by

$$X_\mu(f) = \frac{\partial}{\partial x^\mu}(f).$$

We usually call X_μ simply $\partial/\partial x^\mu$ or ∂_μ , such that all vectors y can be expressed, using the Einstein summation convention, in the form

$$y = y^\mu \partial_\mu, \quad (1.20)$$

where y^μ are the components of y .

The tangent vector to a curve f with coordinates $x^\mu(\tau)$ is then given by the application

$$t(f) = \frac{d}{d\tau} f[x^\mu(\tau)] = \frac{\partial f}{\partial x^\mu} \frac{dx^\mu(\tau)}{d\tau},$$

so that $t = t^\mu \partial_\mu$ with $t^\mu = dx^\mu(\tau)/d\tau$. In a coordinate transformation of the form

$$x^\mu \rightarrow x'^\mu = x'^\mu(x^\nu), \quad \left| \frac{\partial x'^\mu}{\partial x^\nu} \right| \neq 0, \quad (1.21)$$

we have, for all $f \in F$, $\partial_\nu f = (\partial_\nu x'^\mu)(\partial'_\mu f)$. We thus deduce that the elements of the basis are transformed according to

$$\partial_\mu \rightarrow \partial'_\mu = \frac{\partial x^\nu}{\partial x'^\mu} \partial_\nu, \quad (1.22)$$

so that the components of all vectors y transform according to

$$y^\mu \rightarrow y'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} y^\nu, \quad (1.23)$$

which is the standard formula for the coordinate transformation that we have used previously.

1.2.1.3 Tensors

The notion of tensor is a generalization of that of a vector, when one considers quantities that have multilinear dependencies under an infinitesimal transformation.

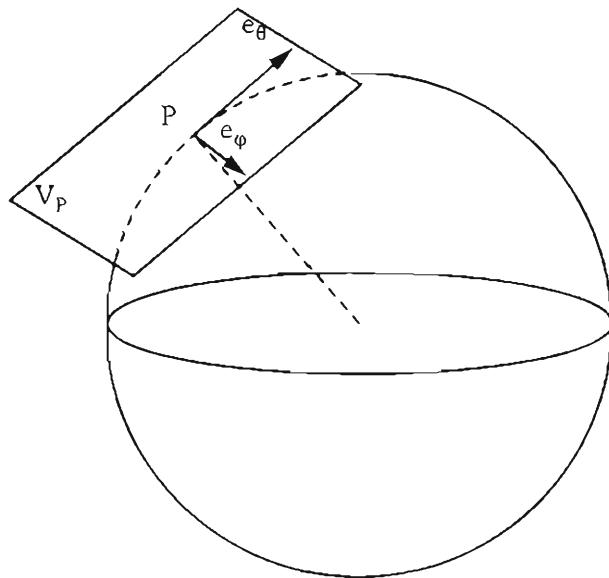


Fig. 1.4 The tangent space of a two-dimensional sphere is a plane whose basis is given by $e_\theta = \partial_\theta$ and $e_\varphi = \partial_\varphi$.

Given a vector space V_p , one can construct a dual space V_p^* composed of the linear maps from V_p to \mathbb{R} . This new vector space has a basis y^μ that can be defined by specifying their action on the elements of a basis y_ν of V_p

$$y^{\mu*}(y_\nu) = \delta_\nu^\mu.$$

In particular, dx^ν represents the basis of V_p^* associated with the basis ∂_μ of V_p . dx^ν should therefore satisfy

$$dx^\nu(\partial_\mu) = \delta_\mu^\nu$$

and it associates any vector y to its component y^ν defined by $dx^\nu(y) = y^\nu$. In a coordinate transformation, dx^μ thus transforms as the components of a vector

$$dx^\mu \rightarrow dx'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} dx^\nu. \quad (1.24)$$

A tensor T of rank (r, q) is a multilinear map from $(V_p)^r \times (V_p^*)^q$ to \mathbb{R} such that

$$T = T_{\nu_1 \dots \nu_q}^{\mu_1 \dots \mu_r} \partial_{\mu_1} \otimes \dots \otimes \partial_{\mu_r} \otimes dx^{\nu_1} \otimes \dots \otimes dx^{\nu_q}, \quad (1.25)$$

where \otimes is the tensor product. Such an object is said to be r times contravariant and q times covariant. In particular, a tensor of rank $(0, 0)$ is a scalar whose value is independent of the system of coordinates, a tensor of rank $(1, 0)$ is a vector and a tensor of rank $(0, 1)$ is a 1-form. Using the transformation laws for vectors and 1-forms, we can deduce that in a coordinate transformation, the components of a tensor transform as

$$T'^{\mu_1 \dots \mu_p}_{\nu_1 \dots \nu_q} = \frac{\partial x'^{\mu_1}}{\partial x^{\alpha_1}} \dots \frac{\partial x'^{\mu_p}}{\partial x^{\alpha_p}} \frac{\partial x'^{\beta_1}}{\partial x'^{\nu_1}} \dots \frac{\partial x'^{\beta_q}}{\partial x'^{\nu_q}} T^{\alpha_1 \dots \alpha_p}_{\beta_1 \dots \beta_q}. \quad (1.26)$$

A tangent space is associated to each point of the manifold. Tensors of the same nature defined at different points will therefore act on different spaces. One could, however, introduce the notion of a tensor field by selecting a tensor in a continuous way.

We can now introduce the notion of a metric. Just as in Minkowski space-time in general coordinates, one has to use the metric to define the square of the distance between two neighbouring points. The metric g should therefore be a symmetric and non-degenerate tensor of rank $(0, 2)$. The line-element between two events is then given by

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu, \quad (1.27)$$

where $g_{\mu\nu}$ is the metric tensor; it is symmetric and has hence 10 components. From the previous results, we can deduce that it transforms as

$$g'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu} \frac{\partial x^\beta}{\partial x'^\nu} g_{\alpha\beta}. \quad (1.28)$$

This expression is analogous to the one obtained in the Euclidean or Minkowski cases. In particular, we can deduce that the determinant of the metric $g = \det g_{\mu\nu}$, transforms as

$$g' = \left[\det \left(\frac{\partial x^\alpha}{\partial x'^\mu} \right) \right]^2 g. \quad (1.29)$$

1.2.1.4 Tensor calculus

Tensors are thus a generalization of the notion of vectors. Tensors of the same rank can be added

$$R^{\mu_1 \dots \mu_p}_{\nu_1 \dots \nu_q} = S^{\mu_1 \dots \mu_p}_{\nu_1 \dots \nu_q} + T^{\mu_1 \dots \mu_p}_{\nu_1 \dots \nu_q},$$

and one can multiply tensors of different ranks to obtain a tensor of higher rank

$$R^{\mu_1 \dots \mu_r}_{\nu_1 \dots \nu_s} = S^{\mu_1 \dots \mu_p}_{\nu_1 \dots \nu_q} T^{\mu_{p+1} \dots \mu_r}_{\nu_{q+1} \dots \nu_s}.$$

The indices can be contracted to form a tensor whose rank is lowered by 2

$$R^{\mu_2 \dots \mu_p}_{\nu_2 \dots \nu_q} = S^{\mu_1 \mu_2 \dots \mu_p}_{\nu_1 \nu_2 \dots \nu_q} = S^{\mu_1 \mu_2 \dots \mu_p}_{\nu_1 \nu_2 \dots \nu_q} \delta^{\nu_1}_{\mu_1}.$$

This operation reduces to sum over the repeated index. This allows us to define a scalar by contracting all the indices. In particular, $T^\mu V_\mu = g_{\mu\nu} V^\mu T^\nu$ represents the scalar product of two vectors. We conclude that $T^{\mu_1 \dots \mu_p}_{\nu_1 \dots \nu_q}$ is a tensor if and only if its contraction with any tensor of the form $R^{\nu_1 \dots \nu_q}_{\mu_1 \dots \mu_p}$ is an invariant.

The metric can be used to raise or lower the indices

$$R^{\mu_1 \dots \mu_p \mu}_{\nu_2 \dots \nu_q} = R^{\mu_1 \dots \mu_p}_{\nu_1 \nu_2 \dots \nu_q} g^{\nu_1 \mu}.$$

In particular, $T^\mu = g^{\mu\nu} T_\nu$.

A tensor is symmetric if it satisfies $T^{\mu\nu} = T^{\nu\mu}$ and is antisymmetric if $T^{\mu\nu} = -T^{\nu\mu}$. From any tensor, we can extract a symmetric and an antisymmetric tensor by

$$T_{(\mu\nu)} = \frac{1}{2} (T_{\mu\nu} + T_{\nu\mu}), \quad T_{[\mu\nu]} = \frac{1}{2} (T_{\mu\nu} - T_{\nu\mu}). \quad (1.30)$$

This can be performed on one or several pairs of indices, or on all the indices. For instance, symmetrizing two indices of a vector of rank 4 gives

$$T_{(\mu_1\alpha\beta\nu)} = \frac{1}{2} (T_{\mu_1\alpha\beta\nu} + T_{\nu\alpha\beta\mu_1}),$$

and a completely symmetric tensor is obtained by

$$T_{(\mu_1\dots\mu_n)} = \frac{1}{n!} (T_{\mu_1\dots\mu_n} + T_{\mu_2\mu_1\dots\mu_n} + \dots + T_{\mu_n\mu_2\dots\mu_1}).$$

In the particular case of a rank 2 tensor, it can always be decomposed as the sum of a traceless symmetric tensor, an antisymmetric tensor and a trace

$$T_{\mu\nu} = T_{[\mu\nu]} + \left(T_{(\mu\nu)} - \frac{1}{4} T^\alpha_\alpha g_{\mu\nu} \right) + \frac{1}{4} T^\alpha_\alpha g_{\mu\nu},$$

in four dimensions.

We note that the completely antisymmetric tensor of rank n is given by $\epsilon_{\mu_1\mu_2\dots\mu_n}$

$$\epsilon_{\mu_1\dots\mu_n} = \epsilon_{[\mu_1\mu_2\dots\mu_n]}, \quad (1.31)$$

with $\epsilon_{1\dots n} = \sqrt{-g}$ and $\epsilon^{1\dots n} = 1/\sqrt{-g}$. For $n = 4$, the antisymmetric tensor has the following property

$$\epsilon_{\alpha\beta\delta\gamma}\epsilon^{\alpha\lambda\nu\mu} = -\delta^\lambda_\beta\delta^\nu_\delta\delta^\mu_\gamma - \delta^\nu_\beta\delta^\mu_\delta\delta^\lambda_\gamma - \delta^\mu_\beta\delta^\lambda_\delta\delta^\nu_\gamma + \delta^\lambda_\beta\delta^\mu_\delta\delta^\nu_\gamma + \delta^\mu_\beta\delta^\nu_\delta\delta^\lambda_\gamma + \delta^\nu_\beta\delta^\lambda_\delta\delta^\mu_\gamma. \quad (1.32)$$

Contracting this identity, we get

$$\epsilon_{\alpha\beta\delta\gamma}\epsilon^{\alpha\beta\nu\mu} = -2 (\delta^\nu_\delta\delta^\mu_\gamma + \delta^\mu_\delta\delta^\nu_\gamma), \quad \epsilon_{\alpha\beta\delta\gamma}\epsilon^{\alpha\beta\delta\mu} = -6\delta^\mu_\gamma, \quad \epsilon_{\alpha\beta\delta\gamma}\epsilon^{\alpha\beta\delta\gamma} = -24. \quad (1.33)$$

1.2.2 Geodesic equations and Christoffel symbols

Using the notions introduced in the previous section, we can go back to the trajectory of a free-falling body moving in a space-time with metric $g_{\mu\nu}$. Such a trajectory is a line of shortest length, such that its equation, $x^\mu(\lambda)$, can be obtained by maximizing the action

$$S = \int ds.$$

However, as already mentioned in the Minkowski case, ds can be negative or null for some trajectories. It is therefore cleverer to maximize the action

$$S = \int L d\lambda, \quad L = g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu, \quad (1.34)$$

with $\dot{x}^\mu = dx^\mu/d\lambda$. This is equivalent to maximizing $\int(ds^2/d\lambda^2)d\lambda$. We can show that extremizing these two actions comes to the same result if $ds \neq 0$, but using the

second one allows us to get the correct result also for null geodesics² ($ds = 0$, see Section 1.4.2).

The Euler–Lagrange equations take the form

$$\frac{\delta L}{\delta x^\alpha} = \frac{d}{d\lambda} \left(\frac{\delta L}{\delta \dot{x}^\alpha} \right).$$

Since

$$\frac{\delta L}{\delta x^\alpha} = \partial_\alpha g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu, \quad \frac{d}{d\lambda} \left(\frac{\delta L}{\delta \dot{x}^\alpha} \right) = \frac{d}{d\lambda} (2g_{\mu\alpha} \ddot{x}^\mu) = 2g_{\mu\alpha} \ddot{x}^\mu + 2\partial_\beta g_{\mu\alpha} \dot{x}^\beta \dot{x}^\mu,$$

we have that

$$g_{\mu\alpha} \ddot{x}^\mu = \frac{1}{2} (\partial_\alpha g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu - \partial_\beta g_{\mu\alpha} \dot{x}^\beta \dot{x}^\mu - \partial_\mu g_{\beta\alpha} \dot{x}^\beta \dot{x}^\mu)$$

where we have symmetrized the term $\partial_\beta g_{\mu\alpha} \dot{x}^\beta \dot{x}^\mu$. Introducing the Christoffel symbols

$$\Gamma_{\lambda\mu\nu} = \frac{1}{2} (\partial_\mu g_{\lambda\nu} + \partial_\nu g_{\mu\lambda} - \partial_\lambda g_{\mu\nu}), \quad (1.35)$$

and

$$\Gamma_{\mu\nu}^\alpha = g^{\lambda\alpha} \Gamma_{\lambda\mu\nu}, \quad (1.36)$$

and contracting our equation with $g^{\alpha\sigma}$, we get the geodesic equation

$$\ddot{x}^\mu + \Gamma_{\nu\lambda}^\mu \dot{x}^\nu \dot{x}^\lambda = 0 \quad (1.37)$$

identical to (1.18) obtained in special relativity. As we have stressed, space-time curvature does not intervene in the equation of the trajectory of a single particle, but it does in the deviation of close trajectories.

From the definition of the Christoffel symbols (1.35), we have, since the metric is symmetric,

$$\Gamma_{\mu\nu}^\alpha = \Gamma_{\nu\mu}^\alpha, \quad (1.38)$$

as well as

$$\Gamma_{\alpha\mu\nu} = \Gamma_{\alpha\nu\mu}. \quad (1.39)$$

Note that in four dimensions there are $4 \times 10 = 40$ Christoffel symbols, which is precisely the number of partial derivatives of the metric $\partial_\alpha g_{\mu\nu}$. We can hence inverse the relation (1.35) to obtain, after some algebra,

$$\partial_\alpha g_{\mu\nu} = g_{\lambda\mu} \Gamma_{\alpha\nu}^\lambda + g_{\lambda\nu} \Gamma_{\alpha\mu}^\lambda. \quad (1.40)$$

Using the geodesic equation, we can convince ourselves that the Christoffel symbols are not tensors since they transform as

$$\Gamma'{}^\alpha_{\mu\nu} = \frac{\partial x'^\alpha}{\partial x^\beta} \frac{\partial x^\sigma}{\partial x'^\mu} \frac{\partial x^\rho}{\partial x'^\nu} \Gamma^\beta_{\sigma\rho} - \frac{\partial^2 x'^\alpha}{\partial x^\sigma \partial x^\rho} \frac{\partial x^\sigma}{\partial x'^\mu} \frac{\partial x^\rho}{\partial x'^\nu} \quad (1.41)$$

under a coordinate transformation.

²Light-like geodesics are also called *isotropic geodesics*; we will avoid this terminology, which can be confusing. We will call them either *null geodesics* or *light-like geodesics* with no distinction.

To finish, we consider the determinant of the metric

$$g = \det g_{\mu\nu} = \pi^{\alpha\beta\gamma\delta}_{0123} g_{\alpha 0} g_{\beta 1} g_{\gamma 2} g_{\delta 3},$$

where $\pi^{\alpha\beta\gamma\delta}_{\mu\nu\rho\sigma} = +1$ if $(\alpha\beta\gamma\delta)$ is an even permutation of $(\mu\nu\rho\sigma)$, -1 if it is an odd permutation and 0 otherwise. We thus obtain the derivative of the determinant

$$\partial_\alpha g = g g^{\mu\nu} \partial_\alpha g_{\mu\nu}. \quad (1.42)$$

Another way to view this result is to write that $dg = m^{\mu\nu} dg_{\mu\nu}$, where $m^{\mu\nu}$ are the minors of the matrix of the $g^{\mu\nu}$. The inverse of a matrix is $g^{\mu\nu} = (g_{\mu\nu})^{-1} = m^{\mu\nu}/g$, so we can deduce that $\partial_\alpha g = m^{\mu\nu} \partial_\alpha g_{\mu\nu} = gg^{\mu\nu} \partial_\alpha g_{\mu\nu}$. We can then deduce the useful property

$$\Gamma^\mu_{\mu\alpha} = \partial_\alpha (\ln \sqrt{-g}). \quad (1.43)$$

We can also point out another interesting property obtained using the fact that $g^{\mu\nu} g_{\mu\nu} = 4$, and hence that $\partial_\alpha (g^{\mu\nu} g_{\mu\nu}) = 0$, that is

$$g^{\mu\nu} \partial_\alpha g_{\mu\nu} = -g_{\mu\nu} \partial_\alpha g^{\mu\nu}. \quad (1.44)$$

1.2.3 Covariant derivative, parallel transport and Lie derivative

The laws of physics often take the form of partial differential equations. It is thus important to understand how a tensor field can be differentiated in a way that preserves its tensor properties.

1.2.3.1 Covariant derivative

The partial derivative of an arbitrary tensor field

$$\partial_\alpha T^{\mu_1 \dots \mu_p}_{\nu_1 \dots \nu_q} (x^\beta) = T^{\mu_1 \dots \mu_p}_{\nu_1 \dots \nu_q, \alpha}$$

is no longer a tensor, apart from the exceptional case of the derivative of a scalar that is a 1-form. To really understand this, let us consider the derivative of a vector T^μ and how it transforms under a coordinate transformation

$$\frac{\partial T'^\mu}{\partial x'^\alpha} = \frac{\partial}{\partial x'^\alpha} \left(\frac{\partial x'^\mu}{\partial x^\nu} T^\nu \right) = \frac{\partial x'^\mu}{\partial x^\nu} \frac{\partial x^\sigma}{\partial x'^\alpha} \frac{\partial T^\nu}{\partial x^\sigma} + \frac{\partial^2 x'^\mu}{\partial x^\sigma \partial x'^\alpha} \frac{\partial x^\sigma}{\partial x'^\alpha} T^\nu. \quad (1.45)$$

The reason for this is that this operation makes the difference between the value of the tensor at x^α and $x^\alpha + dx^\alpha$ and that at these points, the tensor has different transformation laws. Consider a constant vector of the Euclidean plane, with coordinates $T^x = 1, T^y = 0$. In polar coordinates, its coordinates depend on the position ($T^r = \sin \theta, T^\theta = \cos \theta$) simply because the directions of the coordinate lines change from one point to the other. It follows that $\partial_\theta T^r$ and $\partial_\theta T^\theta$ do not vanish even if T is a constant. What we should define is a derivative such that the components of a constant vector do not change when it is transported from x^α to $x^\alpha + dx^\alpha$.

The covariant derivative is the differentiation operator that parallel transports the tensor from x^α to $x^\alpha + dx^\alpha$ before subtracting its value at x^α . Since this operation is linear, one can always define such a covariant derivative by

$$\nabla_\mu T^\nu = \partial_\mu T^\nu + C_{\mu\alpha}^\nu T^\alpha, \quad \nabla_\mu T_\nu = \partial_\mu T_\nu - C_{\mu\nu}^\alpha T_\alpha, \quad (1.46)$$

where the quantities $C_{\mu\alpha}^\nu$ should be determined. For that, we notice that (1.45) implies that these quantities should transform as (1.41) during a change of coordinate. Working locally in Minkowski space and noting that then $\nabla_\alpha T^\mu = \partial_\alpha T^\mu$, we conclude that

$$C_{\mu\alpha}^\nu = \Gamma_{\mu\alpha}^\nu. \quad (1.47)$$

In the rest of this book, we will denote the covariant derivative either with the symbol ∇_μ or by ; μ in subscript, with no distinction: $\nabla_\mu T^\nu = T^\nu_{;\mu}$. Similarly, we will use the notation $\partial_\mu T^\nu = T^\nu_{,\mu}$ for the ordinary partial derivative. The relation (1.46) can be generalized to an arbitrary tensor

$$\begin{aligned} \nabla_\alpha T_{\nu_1 \dots \nu_q}^{\mu_1 \dots \mu_p} &= \partial_\alpha T_{\nu_1 \dots \nu_q}^{\mu_1 \dots \mu_p} + \Gamma_{\alpha\lambda_1}^{\mu_1} T_{\nu_1 \dots \nu_q}^{\lambda_1 \mu_2 \dots \mu_p} + \dots + \Gamma_{\alpha\lambda_p}^{\mu_p} T_{\nu_1 \dots \nu_q}^{\mu_1 \dots \mu_{p-1} \lambda_p} \\ &\quad - \Gamma_{\alpha\nu_1}^{\lambda_1} T_{\lambda_1 \nu_2 \dots \nu_q}^{\mu_1 \dots \mu_p} - \dots - \Gamma_{\alpha\nu_q}^{\lambda_q} T_{\nu_1 \dots \nu_{q-1} \lambda_q}^{\mu_1 \dots \mu_p}. \end{aligned} \quad (1.48)$$

A consequence of this result is that $\nabla_\alpha g_{\mu\nu} = \partial_\alpha g_{\mu\nu} - \Gamma_{\alpha\mu}^\sigma g_{\sigma\nu} - \Gamma_{\alpha\nu}^\sigma g_{\sigma\mu}$. Using the expressions (1.36) for the Christoffel symbols and (1.40) for the derivative of the metric, we get

$$\nabla_\alpha g_{\mu\nu} = 0. \quad (1.49)$$

To each metric, we can therefore associate a covariant derivative. Note that we can define various covariant derivatives but only one is *compatible* with the metric in the sense that it satisfies the property (1.49).

1.2.3.2 Parallel transport

The covariant derivative allows us to define the notion of parallel transport. A vector is parallel transported if $DT^\mu \equiv dx^\nu \nabla_\nu T^\mu = 0$. We can use this condition to define the notion of parallel transport of a vector along a curve of equation $x^\mu(\tau)$ by $DT^\mu/D\tau = 0$, i.e. by

$$\frac{DT^\mu}{D\tau} = u^\nu \nabla_\nu T^\mu = \frac{dT^\mu}{d\tau} + \Gamma_{\nu\alpha}^\mu T^\alpha u^\nu = 0, \quad (1.50)$$

where $u^\mu = dx^\mu/d\tau$. The example of the sphere presented in Fig. 1.5 shows that the result of this operation depends on which path is followed.

A special case of transport along a curve is the auto-parallel transport, i.e. a curve such that the tangent vector is transported parallel to itself; it should therefore satisfy

$$\frac{Du^\mu}{D\tau} = u^\nu \nabla_\nu u^\mu = 0. \quad (1.51)$$

This is simply the equation of the geodesic (1.37) so that we conclude that the geodesics (curves of shortest path) are therefore auto-parallel.

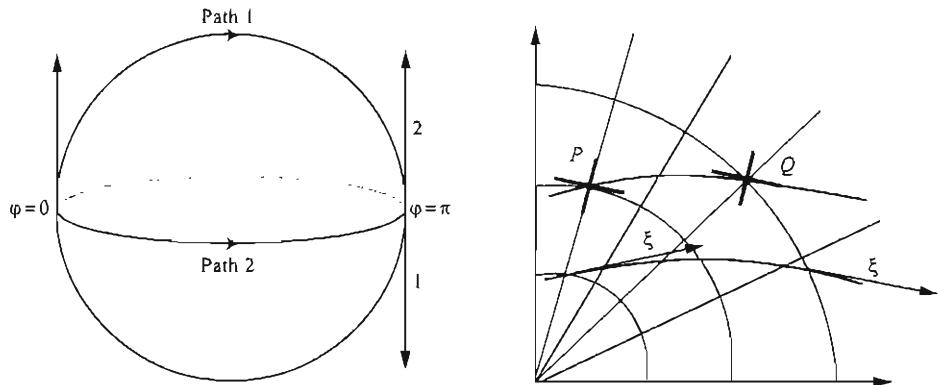


Fig. 1.5 (left): On a 2-dimensional sphere, a vector can be transported parallel to itself along a path 1 corresponding to a great circle until its antipodal point. We can reach the same point following the path 2 along the equator. The two vectors that we get do not coincide. (right): The Lie derivative evaluates the change determined by an observer that goes from a point P to a point Q along the flow curves of the vector field ξ^μ transporting its local frame with it.

1.2.3.3 Differential operators

The covariant derivative allows us to define some mathematical operators such as the divergence and the curl. Using (1.35), we get that

$$\nabla_\mu T^\mu = \partial_\mu T^\mu + \Gamma_{\mu\rho}^\mu T^\rho = \partial_\mu T^\mu + T^\rho \partial_\rho \ln \sqrt{-g},$$

which can be used to express the divergence of any vector field by

$$\nabla_\mu T^\mu = \frac{1}{\sqrt{-g}} \partial_\mu (\sqrt{-g} T^\mu). \quad (1.52)$$

Similarly, we can check that the divergence of any antisymmetric tensor is of the form

$$\nabla_\mu F^{\mu\nu} = \frac{1}{\sqrt{-g}} \partial_\mu (\sqrt{-g} F^{\mu\nu}), \quad (1.53)$$

where we used the fact that the contraction of a symmetric tensor with an antisymmetric one vanishes.

The d'Alembertian operator is defined as $\square = \nabla_\mu \nabla^\mu$ and can be applied to any kind of tensor. In a Minkowski space-time, it reduces to $\nabla_\mu \nabla^\mu = -\partial_t^2 + \Delta$, where $\Delta = \delta^{ij} \partial_i \partial_j$ is the usual three-dimensional Laplacian. Note that the relation (1.52) can be used to express the d'Alembertian of any scalar by

$$\nabla_\mu \nabla^\mu \phi = \frac{1}{\sqrt{-g}} \partial_\mu (\sqrt{-g} g^{\mu\nu} \partial_\nu \phi), \quad (1.54)$$

where we used the fact that $\nabla_\mu \phi = \partial_\mu \phi$. Finally, the analogue of the curl of a 1-form A_μ is given by

$$A_{\nu;\mu} - A_{\mu;\nu} = A_{\nu,\mu} - A_{\mu,\nu}, \quad (1.55)$$

since $\Gamma_{\mu\nu}^\alpha$ is symmetric in $\mu\nu$.

1.2.3.4 Lie derivative

Another way to construct a derivative is to consider the Lie derivative. For that we assume that there exists a vector field ξ^μ and try to define how the coordinates of a vector T^μ will change for an observer that moves along the flow curves of the vector ξ^μ from a point P with coordinates x^α to a point Q with coordinates $x^\alpha + \epsilon \xi^\alpha$ (where ϵ is a small parameter) and that transports its system of coordinates with itself (see Fig. 1.5).

Using the coordinates defined in P at the point Q is equivalent to performing a coordinate transformation in Q defined as $x'^\mu = x^\mu - \epsilon \partial_\nu \xi^\mu$, such that

$$\frac{\partial x'^\mu}{\partial x^\nu} = \delta_\nu^\mu - \epsilon \partial_\nu \xi^\mu,$$

and

$$\frac{\partial x^\nu}{\partial x'^\mu} = \delta_\mu^\nu + \epsilon \partial^\nu \xi_\mu,$$

to leading order in ϵ . The coordinates of the vector T^μ at Q will be given by

$$T'^\mu(Q) = \frac{\partial x'^\mu}{\partial x^\nu} T^\nu(x^\alpha + \epsilon \xi^\alpha) = (\delta_\nu^\mu - \epsilon \partial_\nu \xi^\mu)[T^\nu(P) + \epsilon \xi^\alpha \partial_\alpha T^\nu(P)].$$

The Lie derivative is obtained by comparing this quantity with the vector T^μ at P

$$\mathcal{L}_\xi T^\mu = \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} [T'^\mu(Q) - T^\mu(P)],$$

so that

$$\mathcal{L}_\xi T^\mu = \xi^\nu \partial_\nu T^\mu - T^\nu \partial_\nu \xi^\mu. \quad (1.56)$$

This relation can be generalized to any tensor by using the definition (1.56) and the relevant transformation laws

$$\mathcal{L}_\xi T_{\nu_1 \dots \nu_q}^{\mu_1 \dots \mu_p} = \xi^\alpha \partial_\alpha T_{\nu_1 \dots \nu_q}^{\mu_1 \dots \mu_p} - \sum_{i=1}^p T_{\nu_1 \dots \nu_q}^{\mu_1 \dots \alpha \dots \mu_p} \partial_\alpha \xi^{\mu_i} + \sum_{j=1}^q T_{\nu_1 \dots \beta \dots \nu_q}^{\mu_1 \dots \mu_p} \partial_{\nu_j} \xi^\beta. \quad (1.57)$$

The Lie derivative plays an important role in the study of the symmetries of a Riemannian space. If $\mathcal{L}_\xi T_\nu^\mu = 0$ then $T'_\nu^\mu(x') = T_\nu^\mu(x')$ and the tensor field is invariant under that change of coordinates. In particular, for the metric tensor, using (1.57) for a symmetric rank-2 tensor, the relations (1.40) and (1.38), as well as the definition of the covariant derivative for a vector, we get

$$\mathcal{L}_\xi g_{\mu\nu} = \nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu. \quad (1.58)$$

1.2.4 Curvature

1.2.4.1 Curvature tensor

As seen from Fig. 1.5, the parallel transport of a vector along a closed curve does not bring the vector back to its initial state if the space is curved, contrary to what

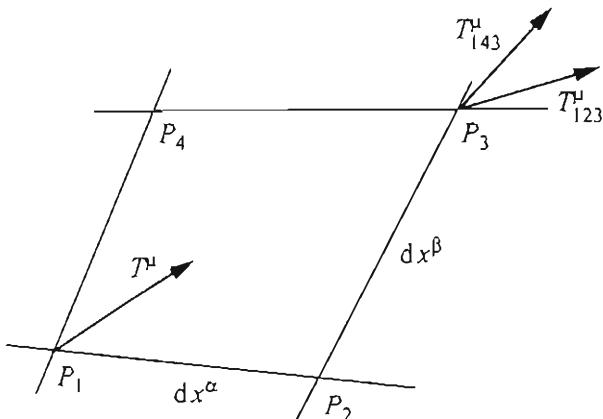


Fig. 1.6 The vector T^μ is parallel transported along the path P_1P_3 passing by either P_2 or P_4 to give, respectively, the vectors T_{123}^μ and T_{143}^μ . These two vectors only coincide if the Riemann tensor vanishes.

happens in a Euclidean space. Just like for the geodesic deflection, this characterizes the curvature of space.

Let us consider the parallel transport of a vector T^μ along the closed path shown in Fig. 1.6. From P_1 to P_2 , the relation (1.50) implies that the variation of the vector T^μ is given by $\delta_{12}T^\mu = dx^\alpha \nabla_\alpha T^\mu = dT^\mu + \Gamma_{\alpha\sigma}^\mu T^\sigma dx^\alpha$. By iterating it, we get that the variation between P_1 and P_3 is given by $\delta_{123}T^\mu = dx^\alpha dx^\beta \nabla_\beta \nabla_\alpha T^\mu = d^2T^\mu + 2\Gamma_{\alpha\nu}^\mu dx^\alpha dT^\nu + (\partial_\beta \Gamma_{\alpha\nu}^\mu + \Gamma_{\beta\nu}^\mu \Gamma_{\alpha\beta}^\nu) dx^\alpha dx^\beta T^\nu$. The same variation passing through P_4 is given by $\delta_{143}T^\mu = dx^\alpha dx^\beta \nabla_\alpha \nabla_\beta T^\mu$. On comparing these two expressions, only the terms proportional to $dx^\alpha dx^\beta$ are different and we get that

$$(\nabla_\alpha \nabla_\beta - \nabla_\beta \nabla_\alpha) T^\mu = R^\mu_{\nu\alpha\beta} T^\nu, \quad (1.59)$$

where the Riemann tensor is given by

$$R^\mu_{\nu\alpha\beta} = \partial_\alpha \Gamma_{\nu\beta}^\mu - \partial_\beta \Gamma_{\nu\alpha}^\mu + \Gamma_{\sigma\alpha}^\mu \Gamma_{\nu\beta}^\sigma - \Gamma_{\sigma\beta}^\mu \Gamma_{\nu\alpha}^\sigma, \quad R_{\sigma\nu\alpha\beta} = g_{\sigma\mu} R^\mu_{\nu\alpha\beta}. \quad (1.60)$$

The definition (1.60) allows one to conclude that the Riemann tensor is antisymmetric in the permutation of the first or the second pair of indices

$$R_{\alpha\mu\nu\sigma} = -R_{\mu\alpha\nu\sigma} = -R_{\alpha\mu\sigma\nu}, \quad (1.61)$$

and that it is symmetric in the permutation of the two pairs

$$R_{\alpha\mu\nu\sigma} = R_{\nu\sigma\alpha\mu}. \quad (1.62)$$

We can also show that it satisfies the identity

$$3R_{[\mu\nu\sigma]} = R_{\alpha\mu\nu\sigma} + R_{\alpha\nu\sigma\mu} + R_{\alpha\sigma\mu\nu} = 0. \quad (1.63)$$

The two first symmetry properties imply that the Riemann tensor has as many components as a 6×6 symmetric matrix, i.e. 21 components. The last relation (1.63) is

independent of the first ones and thus adds one extra constraint to the number of components. The Riemann tensor therefore has 20 independent components.

Finally, the derivatives of the Riemann tensor enjoy a cyclic symmetry property that is

$$3R_{\alpha\mu|\nu\sigma;\beta} = R_{\alpha\mu\nu\sigma,\beta} + R_{\alpha\mu\sigma\beta,\nu} + R_{\alpha\mu\beta\nu,\sigma} = 0, \quad (1.64)$$

called the Bianchi identity.

1.2.4.2 Ricci and Einstein tensors

The symmetry properties of the Riemann tensor imply that we can construct only one tensor by contraction of two indices. This is the Ricci tensor, which is symmetric and defined by

$$R_{\mu\nu} = R^\alpha{}_{\mu\alpha\nu} = -R^\alpha{}_{\mu\nu\alpha}. \quad (1.65)$$

The Ricci tensor has 10 independent components. Its trace defines a scalar, the scalar curvature, also known as the Ricci scalar

$$R = R_{\mu\nu}g^{\mu\nu}. \quad (1.66)$$

Contracting the Bianchi identities (1.64), respectively, by $g^{\alpha\beta}$ and $g^{\alpha\beta}g^{\mu\sigma}$ implies that

$$R^\alpha{}_{\mu\nu\sigma;\alpha} - R_{\mu\sigma;\nu} + R_{\mu\nu;\sigma} = 0, \quad R^\alpha{}_{\nu;\alpha} - R_{;\nu} + R^\alpha{}_{\nu;\alpha} = 0.$$

We can combine these two equations to construct a conserved tensor for the covariant derivative

$$G^{\mu\nu}{}_{;\nu} = 0, \quad G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}.$$

(1.67)

This tensor $G_{\mu\nu}$ is called the Einstein tensor.

1.2.4.3 Killing vector fields

A vector that satisfies the Killing equation

$$\mathcal{L}_\xi g_{\mu\nu} = 0$$

is called a *Killing vector*; it keeps the metric invariant and therefore corresponds to a space-time symmetry. From the relation (1.59), we conclude that (1.58) implies that $\nabla_\alpha \nabla_\beta \xi_\nu = R^\mu{}_{\alpha\beta\nu\xi_\mu}$. This shows that from the value of ξ_ν and $\nabla_\mu \xi_\nu$ at a given point one can determine the Killing vector field uniquely. One should then specify N values for ξ_ν and $(N-1)/2$ values for $\nabla_\mu \xi_\nu$, keeping in mind it is antisymmetric, so that there are at most $N(N+1)/2$ linearly independent Killing vector fields. For $N=4$, this is 10 Killing vectors, which is precisely the dimension of the Poincaré group of Minkowski space.

Indeed, this maximum number of Killing vector fields is not always reached in an arbitrary space-time since the Killing equations are not necessarily integrable for any choice of initial conditions; most space-times have no symmetries and no Killing vector fields!

A space-time enjoying the maximum number of Killing vector fields is called a maximally symmetric space-time. It can be shown from the Killing equations and (1.59) that the Riemann tensor must then satisfy (see, e.g., Ref. [2] for a demonstration)

$$(N - 1)R^\mu_{\alpha\beta\nu} = R_{\alpha\nu}\delta_\beta^\mu - R_{\alpha\beta}\delta_\nu^\mu,$$

so that

$$NR_\nu^\mu = \frac{R}{N(N - 1)} (R_{\alpha\nu}\delta_\beta^\mu - R_{\alpha\beta}\delta_\nu^\mu),$$

the Ricci scalar, R , being a constant. Maximally symmetric spaces are thus spaces of constant curvature. For $N = 4$, there are three maximally symmetric space-times: Minkowski, de Sitter and anti-de Sitter. See also the discussion in Section 3.1.3 of Chapter 3.

1.2.5 Covariant approach

We consider a family of time-like geodesics representing the worldlines of typical observers of the Universe (the fundamental observers). In the case of a fluid, we can think of this flow of geodesics as the set of worldline fluid elements. Each observer has a tangent vector to its worldline, $u^\mu = dx^\mu/d\tau$. This vector is time-like and satisfies³

$$u_\mu u^\mu = -1.$$

A more detailed description of this covariant approach can be found in Ref. [9].

1.2.5.1 3+1 decomposition

With the vector u^μ one can define two projection tensors

$$U_\nu^\mu \equiv -u^\mu u_\nu, \quad \gamma_{\mu\nu} = g_{\mu\nu} + u_\mu u_\nu, \quad (1.68)$$

along and perpendicular to the vector u^μ , respectively. We have

$$U_\nu^\mu U_\alpha^\nu = U_\alpha^\mu, \quad U_\mu^\mu = 1, \quad U_\nu^\mu u^\nu = u^\mu,$$

and

$$\gamma_\nu^\mu \gamma_\alpha^\nu = \gamma_\alpha^\mu, \quad \gamma_\mu^\mu = 3, \quad \gamma_\nu^\mu u^\nu = 0.$$

These two projectors allow one to define a natural notion of space and time for an observer O since the line-element takes the form

$$ds^2 = -(u_\mu dx^\mu)^2 + \gamma_{\mu\nu} dx^\mu dx^\nu. \quad (1.69)$$

Any vector v^μ can be decomposed into ‘time’ and ‘space’ parts, relative to the vector u^μ as

$$v^\mu = v^\nu U_\nu^\mu + v_\perp^\mu, \quad v_\perp^\mu = v^\nu \gamma_\nu^\mu.$$

In particular, a particle of mass m has momentum

³We now work in a unit system such that $c = 1$. This equation would otherwise take the form $u_\mu u^\mu = -c^2$.

$$p^\mu = mv^\mu, \quad p^\mu p_\mu = -m^2.$$

The observer O will measure the energy of this particle to be

$$E = -p^\mu u_\mu, \quad (1.70)$$

so that $E^2 - p_{\perp\mu} p_\perp^\mu = m^2$. Note that if this particle does not follow a geodesic, then its equation of motion is

$$ma^\mu = mv^\nu \nabla_\nu v^\mu = F^\mu, \quad F^\mu v_\mu = 0, \quad (1.71)$$

where F^μ represents any non-gravitational force applied on the particle. Here, the last equality stems from the fact that $v^\mu v_\mu = -1$.

1.2.5.2 Spatial and time derivatives

We can now define a derivative along a worldline, tangent to u^μ by

$$\dot{T}_{\mu\nu\dots}^{\alpha\beta\dots} = u^\lambda \nabla_\lambda T_{\mu\nu\dots}^{\alpha\beta\dots}.$$

It corresponds to a derivative with respect to the proper time. We can also define a derivative in the hypersurface orthogonal to u^μ by

$$\bar{\nabla}_\lambda T_{\mu\nu\dots}^{\alpha\beta\dots} = \gamma_\lambda^{\lambda_1} (\gamma_{\alpha_1}^\alpha \gamma_{\beta_1}^\beta \dots) (\gamma_{\mu_1}^{\mu_1} \gamma_{\nu_1}^{\nu_1} \dots) \nabla_{\lambda_1} T_{\mu_1 \nu_1 \dots}^{\alpha_1 \beta_1 \dots}.$$

This derivative is a three-dimensional covariant derivative associated to the spatial metric (we can indeed check that $\bar{\nabla}_\alpha \gamma_{\mu\nu} = 0$). However, note that for an arbitrary function ρ , $\bar{\nabla}_{[\mu} \bar{\nabla}_{\nu]} \rho = u_{[\mu;\nu]} \dot{\rho}$ (recall u_μ is geodesic). Thus, this term vanishes only if $u_{[\mu;\nu]} = 0$. The derivative $\bar{\nabla}$ is well defined only under this condition. We may note that in particular $\bar{\nabla}_\mu T^\mu = \nabla_\mu T^\mu + u_\mu \dot{T}^\mu$.

1.2.5.3 Distortion of the geodesic flow

We consider a vector flow u^μ describing, for instance, a flow of matter. The covariant derivative of the vector field u^μ can be decomposed as

$$\nabla_\mu u_\nu = -u_\mu \dot{u}_\nu + \frac{1}{3} \Theta \gamma_{\mu\nu} + \sigma_{\mu\nu} + \omega_{\mu\nu}. \quad (1.72)$$

The trace $\Theta = \nabla_\mu u^\mu$ is called *expansion* and describes the isotropic distortion, $\sigma_{\mu\nu}$ is a traceless symmetric tensor called *shear* ($\sigma_{\mu\nu} u^\mu = 0$, $\sigma_\mu^\mu = 0$) and describes the distortion of the flow of matter, $\omega_{\mu\nu}$ is an antisymmetric tensor called *vorticity* that describes the rotation of the matter flow ($\omega_{\mu\nu} u^\mu = 0$). The vorticity vector is defined by

$$\omega^\sigma = \frac{1}{2} u_\alpha \varepsilon^{\alpha\sigma\mu\nu} \omega_{\mu\nu} \iff \omega_{\mu\nu} = u^\beta \varepsilon_{\beta\sigma\mu\nu} \omega^\sigma,$$

and satisfies by construction $\omega^\mu u_\mu = \omega_{\mu\nu} \omega^\mu = 0$. The shear and vorticity amplitude are, respectively, defined as

$$\sigma^2 = \frac{1}{2} \sigma_{\mu\nu} \sigma^{\mu\nu}, \quad \omega^2 = \frac{1}{2} \omega_{\mu\nu} \omega^{\mu\nu}.$$

\dot{u}_ν is the *acceleration* and vanishes if u^μ follows a geodesic.

It will be useful to introduce a characteristic scale $S(\tau)$ defined by

$$\frac{\dot{S}}{S} = \frac{1}{3}\Theta,$$

which determines S along each worldline up to an overall factor. The volume of each element of the fluid varies as S^3 .

1.2.5.4 Raychaudhuri equation

We now calculate $\dot{\Theta} = u^\nu \nabla_\nu (\nabla_\mu u^\mu)$ using the decomposition (1.72) as well as (1.59) to commute the covariant derivatives. This leads to the Raychaudhuri equation

$$\dot{\Theta} = -\frac{1}{3}\Theta^2 + \dot{u}^\mu \dot{u}_\mu + 2(\omega^2 - \sigma^2) + \bar{\nabla}_\mu \dot{u}^\mu - R_{\mu\nu} u^\mu u^\nu. \quad (1.73)$$

This equation is at the heart of the study of gravitational collapse. Two other equations describing the projection of the vorticity and of the shear can be obtained in a similar way (see Ref. [9]).

This equation has an important consequence. Let us suppose that the vorticity vanishes ($\omega = 0$) and that $\dot{u}_\mu = 0$ for all times. It then follows that if $R_{\mu\nu} u^\mu u^\nu > 0$ for all times, we have that $\dot{\Theta} + \Theta^2/3 \leq 0$ so that

$$\frac{1}{\Theta(\tau)} \geq \frac{1}{\Theta_0} + \frac{\tau}{3}.$$

- If $\Theta_0 < 0$, i.e. if it is a converging flow, then at a given time $\tau \leq -3/\Theta_0$, $1/\Theta$ must go through zero and Θ should therefore diverge. The geodesic flow is hence pathological. This signals the presence of a caustic or of a singularity.

- If $\Theta_0 > 0$ at a given time, we can express the Raychaudhuri equation as a function of S to obtain that $\ddot{S}/S \leq 0$. There must therefore exist a previous time $\tau_0 < 3/\Theta_0$ such that $S \rightarrow 0$ when $\tau \rightarrow \tau_0$. This translates into the existence of a singularity in the past.

This is a special case of the theorems on space-time singularities that is discussed in detail in Ref. [10]. We give in the following table the definition of several energy conditions that will appear throughout this book.

| Name | Definition | Meaning |
|----------|---|---|
| null | $G_{\mu\nu} k^\mu k^\nu \geq 0$ $\forall k^\mu, k_\mu k^\mu = 0$ | Light is focused by matter |
| causal | $G_{\mu\nu} u^\nu G^{\mu\rho} u_\rho \leq 0$ $\forall u^\mu, u^\mu u_\mu = -1$ | Energy does not flow faster than the speed of light |
| weak | $G_{\mu\nu} u^\mu u^\nu \geq 0$ $\forall u^\mu, u^\mu u_\mu = -1$ | The energy density is positive for any observer |
| dominant | causal and strong | |
| strong | $R_{\mu\nu} u^\mu u^\nu \geq 0$ | $\ddot{S} < 0$ |

1.2.5.5 Case of a vanishing vorticity

If $\omega = 0$ for all times, then the hypersurfaces that are locally perpendicular to the flow u^μ can be globally extended. $\gamma_{\mu\nu}$ is then the spatial metric (i.e. three-dimensional Riemannian metric) induced by $g_{\mu\nu}$ on Σ and $\bar{\nabla}$ is the covariant derivative associated with $\gamma_{\mu\nu}$, $\bar{\nabla}_\alpha \gamma_{\mu\nu} = 0$ (see Fig. 1.7). The ‘bending’ of Σ in the four-dimensional space-time \mathcal{M} can be characterized by the change of direction of the normal u^μ when moved on Σ . This is associated to the *extrinsic curvature tensor*

$$K_{\mu\nu} \equiv \gamma_\mu^\alpha \nabla_\alpha u_\nu = \frac{1}{2} \mathcal{L}_u \gamma_{\mu\nu}. \quad (1.74)$$

Using (1.72), it reduces to

$$K_{\mu\nu} = \frac{1}{3} \Theta \gamma_{\mu\nu} + \sigma_{\mu\nu}.$$

A Riemann⁴ tensor, ${}^{(3)}R_{\alpha\beta\mu\nu}$ can be associated to $\bar{\nabla}$ from (1.59). Starting from the equality $\bar{\nabla}_\alpha \bar{\nabla}_\beta v_\nu = \gamma_\alpha^{\alpha'} \gamma_\beta^{\beta'} \gamma_\nu^{\nu'} \nabla_{\alpha'} (\gamma_{\beta'}^\rho \gamma_{\nu'}^\sigma \nabla_\rho v_\sigma)$ for a 1-form v_σ on Σ (i.e. $v_\sigma u^\sigma = 0$), we get that

$$\bar{\nabla}_\alpha \bar{\nabla}_\beta v_\nu = \gamma_\alpha^{\alpha'} \gamma_\beta^{\beta'} \gamma_\nu^{\nu'} \nabla_{\alpha'} \nabla_{\beta'} v_\sigma + \gamma_\nu^\sigma K_{\alpha\beta} u^\rho \nabla_\rho v_\sigma + \gamma_\beta^\rho K_{\alpha\nu} u^\sigma \nabla_\rho v_\sigma,$$

where we have used the fact that $\gamma_\alpha^\sigma \gamma_\nu^\delta \nabla_\sigma \gamma_\delta^\mu = K_{\alpha\nu} u^\mu$. When taking the antisymmetric part (over α and β), the second term vanishes so that we finally obtain the relation

$${}^{(3)}R_{\alpha\beta\mu}^\nu = \gamma_\alpha^{\alpha'} \gamma_\beta^{\beta'} \gamma_\mu^{\mu'} \gamma_\nu^{\nu'} R_{\alpha'\beta'\mu'}^{\nu'} - K_{\alpha\mu} K_{\beta}^\nu + K_{\beta\mu} K_{\alpha}^\nu, \quad (1.75)$$

where we have used $\gamma_\beta^\rho u^\sigma \nabla_\rho v_\sigma = -K_\beta^\sigma v_\sigma$ to express the two last terms. This is the Gauss relation. Contracted on the two indices β and ν it gives

$${}^{(3)}R_{\alpha\mu} = \gamma_\alpha^{\alpha'} \gamma_\mu^{\mu'} R_{\alpha'\mu'} + \gamma_\alpha^{\alpha'} \gamma_\mu^{\mu'} u^\beta u_\nu R_{\alpha'\beta\mu'}^\nu - K K_{\alpha\mu} + K_{\beta\mu} K_\beta^\nu,$$

and taking the trace, we get the *scalar Gauss relation*

$${}^{(3)}R + K^2 - K_{\alpha\beta} K^{\alpha\beta} = R + 2R_{\mu\nu} u^\mu u^\nu = 2G_{\mu\nu} u^\mu u^\nu. \quad (1.76)$$

It relates the intrinsic curvature of Σ to the Ricci scalar of \mathcal{M} and the extrinsic curvature. To finish, the relation (1.59) for the vector u^μ projected on Σ leads to the *Codazzi relation*

$$\gamma_\rho^\mu n^\sigma \gamma_\alpha^{\alpha'} \gamma_\beta^{\beta'} R_{\sigma\alpha'\beta'}^\rho = \bar{\nabla}_\beta K_\alpha^\mu - \bar{\nabla}_\alpha K_\beta^\mu,$$

which gives, after contraction, the *contracted Codazzi relation*

$$\gamma_\alpha^\mu n^\nu R_{\mu\nu} = \bar{\nabla}_\mu K_\alpha^\mu - \bar{\nabla}_\alpha K. \quad (1.77)$$

⁴In three dimensions, the Riemann tensor is fully determined from the Ricci tensor and the scalar curvature as

$$R^i_{jkl} = \delta_k^i R_{jl} - \delta_l^i R_{jk} + \gamma_{jl} R_k^i - \gamma_{jk} R_l^i + \frac{1}{2} R(\delta_l^i \gamma_{jk} - \delta_k^i \gamma_{jl}).$$

Using the expression of the extrinsic curvature, the scalar Gauss relation (1.76) implies that the curvature of Σ is given by

$${}^{(3)}R = -\frac{2}{3}\Theta^2 + 2\sigma^2 + 2G_{\mu\nu}u^\mu u^\nu, \quad (1.78)$$

which is valid for all space-times such that $\omega = 0$. Such a flow is called *irrotational*.

1.2.5.6 Decomposition of rank 2 tensors

The existence of the vector field u^μ allows one to decompose any symmetric of rank 2 tensor in a unique way, in the form

$$T_{\mu\nu} = Au_\mu u_\nu + 2q_{(\mu}u_{\nu)} + B\gamma_{\mu\nu} + \pi_{\mu\nu}, \quad (1.79)$$

with

$$q_\mu u^\mu = 0, \quad \pi_{\mu\nu}u^\mu = 0, \quad \pi_{(\mu\nu)} = \pi_{\mu\nu}, \quad \pi_{\mu\nu}u^\mu = 0, \quad \pi_{\mu\nu}\gamma^{\mu\nu} = 0.$$

In particular, we have

$$A = T_{\mu\nu}u^\mu u^\nu, \quad 3B = T_{\mu\nu}\gamma^{\mu\nu}.$$

1.3 Equations of motion

1.3.1 Einstein's equations

Similarly to electromagnetism, we would like to formulate second-order equations for $g_{\mu\nu}$. For that, we need a Lagrangian that depends on the metric and its first derivatives only.

The Lagrangian proposed by Einstein and Hilbert reduces to the scalar curvature. Other choices are possible, such as any function of the scalar curvature or terms involving other invariants such as the Gauss–Bonnet term, or even more complicated theories of ‘scalar-tensor’ type; all these cases are discussed and studied in Chapter 10. The ‘simplest’ choice is in perfect agreement with all observations and defines the theory of general relativity. Any other choice will be considered as an extension of the standard case.

Thus, we focus on the action

$$S = \frac{1}{2\kappa} \int (R - 2\Lambda)\sqrt{-g} d^4x + \int \mathcal{L}_{\text{mat}}\sqrt{-g} d^4x, \quad (1.80)$$

where κ is a coefficient to be determined by requiring that the theory reduces to Newtonian gravity in the weak-field limit; this is the only free parameter of the theory. \mathcal{L}_{mat} is the Lagrangian for the matter fields and Λ is a constant called the *cosmological constant*. We will now present the variation of these two terms.

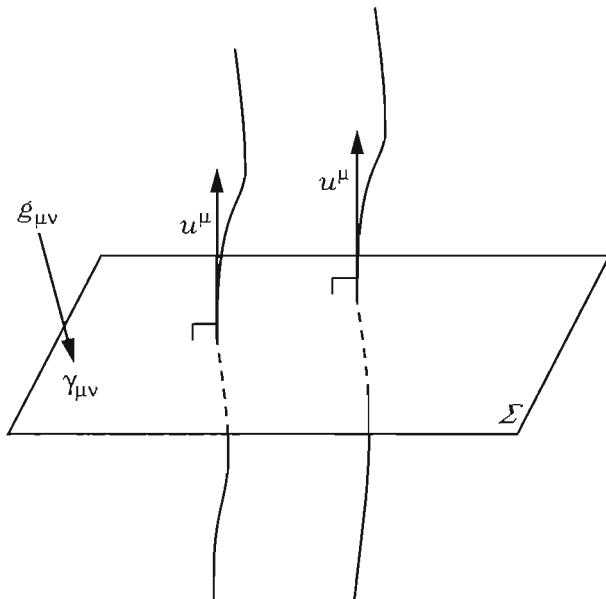


Fig. 1.7 A vector flow u^μ defines an orthogonal surface at each point of the flow. We can then define a global notion of space and time for this flow if and only if its vorticity vanishes.

1.3.1.1 Einstein–Hilbert action

We start by varying the Einstein–Hilbert action

$$\delta \int R \sqrt{-g} d^4x = \int \delta (R_{\mu\nu} g^{\mu\nu} \sqrt{-g}) d^4x,$$

where we have just used the definition of the scalar curvature. Expanding this expression, we get

$$\int \delta (R \sqrt{-g}) d^4x = \int R \delta \sqrt{-g} d^4x + \int g^{\mu\nu} \sqrt{-g} \delta R_{\mu\nu} d^4x + \int R_{\mu\nu} \sqrt{-g} \delta g^{\mu\nu} d^4x.$$

To evaluate the first term, we should vary the metric determinant. Using the result (1.42), we get

$$\delta g = g^{\mu\nu} g \delta g_{\mu\nu} = -g_{\mu\nu} g \delta g^{\mu\nu}, \quad (1.81)$$

which leads to the conclusion

$$\int R \delta \sqrt{-g} d^4x = -\frac{1}{2} \int R \sqrt{-g} g_{\mu\nu} \delta g^{\mu\nu} d^4x.$$

The second term is more complicated to evaluate. For that, we work in a locally inertial coordinate system, i.e. where $g_{\mu\nu} = \eta_{\mu\nu}$ and $\partial_\gamma g_{\mu\nu} = 0$ from which we deduce that

$\Gamma_{\mu\nu}^\alpha = 0$ (but note that the derivatives of the Christoffel symbols do not vanish). We hence obtain

$$\delta R_{\mu\nu} = (\delta \Gamma_{\mu\nu}^\sigma)_{,\sigma} - (\delta \Gamma_{\sigma\nu}^\sigma)_{,\mu} = (\delta \Gamma_{\mu\nu}^\sigma)_{,\sigma} - (\delta \Gamma_{\sigma\nu}^\sigma)_{,\mu}.$$

To go from the partial derivative to the covariant derivative, we used the fact that the result is a tensor, and hence it should be the same in any frame. We can hence conclude that

$$\int_V g^{\mu\nu} \sqrt{-g} \delta R_{\mu\nu} d^4x = \int_V \sqrt{-g} \left[(g^{\mu\nu} \delta \Gamma_{\mu\nu}^\sigma)_{,\sigma} - (g^{\mu\nu} \delta \Gamma_{\sigma\nu}^\sigma)_{,\mu} \right] d^4x,$$

on any volume V . The integrand is therefore of the form $\sqrt{-g} I_{;\mu}^\mu$, which can be re-expressed, using the property (1.52), as $(\sqrt{-g} I^\mu)_{,\mu}$. This is therefore the integral of a total derivative, which can be reduced to an integral on the boundary of the volume V so that it does not contribute to the equations of motion.

We therefore obtain

$$\delta \int R \sqrt{-g} d^4x = \int \sqrt{-g} G_{\mu\nu} \delta g^{\mu\nu} d^4x, \quad (1.82)$$

where the Einstein tensor $G_{\mu\nu}$ is defined in (1.67). The variation of the term in Λ is trivial and gives

$$\int 2\Lambda \delta \sqrt{-g} d^4x = - \int \Lambda \sqrt{-g} g_{\mu\nu} \delta g^{\mu\nu} d^4x.$$

1.3.1.2 Stress-energy tensor

The variation of the matter action with respect to the metric is given by

$$\delta \int \mathcal{L}_{\text{mat}} \sqrt{-g} d^4x = \int \frac{\delta(\mathcal{L}_{\text{mat}} \sqrt{-g})}{\sqrt{-g} \delta g^{\mu\nu}} \delta g^{\mu\nu} \sqrt{-g} d^4x,$$

which we may write as

$$\delta \int \mathcal{L}_{\text{mat}} \sqrt{-g} d^4x = -\frac{1}{2} \int \sqrt{-g} T_{\mu\nu} \delta g^{\mu\nu} d^4x, \quad (1.83)$$

where the energy-momentum tensor $T_{\mu\nu}$, defined by

$$T_{\mu\nu} = -\frac{2}{\sqrt{-g}} \frac{\delta(\mathcal{L}_{\text{mat}} \sqrt{-g})}{\delta g^{\mu\nu}} = g_{\mu\nu} \mathcal{L}_{\text{mat}} - 2 \frac{\delta \mathcal{L}_{\text{mat}}}{\delta g^{\mu\nu}}, \quad (1.84)$$

is a rank 2 symmetric tensor.

1.3.1.3 Einstein's equations

Using the results of the two previous sections, we obtain the general form of Einstein's equations

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = \kappa T_{\mu\nu}. \quad (1.85)$$

Note that there is another way to vary the Lagrangian, known as the Palatini method, in which the metric and the Christoffel symbols are considered as independent.

Extremizing the action with respect to the two fields then gives Einstein's equations and the condition that $\nabla_\mu g_{\alpha\beta} = 0$, which is equivalent to the definition (1.35).

1.3.2 Conservation equations

The Bianchi identities (1.64) impose $G_{;\nu}^{\mu\nu} = 0$, which implies, using Einstein's equations and the fact that $g_{\mu\nu;\alpha} = 0$, that

$$\boxed{T^{\mu\nu}_{;\nu} = 0.} \quad (1.86)$$

This equation is satisfied by the total energy-momentum tensor (i.e. of all matter components). To go further, one should specify the form of the matter Lagrangian or of its energy-momentum tensor.

1.3.2.1 Covariant form

The most general form for the energy-momentum tensor is the one given in (1.79) valid for any symmetric energy-momentum rank 2 tensor,

$$T_{\mu\nu} = \rho u_\mu u_\nu + P \gamma_{\mu\nu} + 2q_{(\mu} u_{\nu)} + \pi_{\mu\nu},$$

in which $\rho = T_{\mu\nu} u^\mu u^\nu$ is the energy density measured by an observer at rest with the fluid, $P = T_{\mu\nu} \gamma^{\mu\nu}/3$ is the pressure, $q^\mu = -T_{\alpha\beta} u^\alpha \gamma^{\beta\mu}$ is the energy flux relative to u^μ and $\pi_{\mu\nu}$ is the anisotropic pressure tensor ($\pi_{\mu\nu} u^\mu = \pi^\mu_\mu = 0$).

The conservation equation $\nabla_\mu T^{\mu\nu} = 0$ can be projected along u^μ and γ_ν^λ , respectively, to obtain the two conservation equations

$$u_\nu \nabla_\mu T^{\mu\nu} = 0, \quad \gamma_\nu^\lambda \nabla_\mu T^{\mu\nu} = 0.$$

Using the decomposition (1.72), we obtain $-u_\nu \nabla_\mu T^{\mu\nu} = u^\mu \nabla_\mu \rho + \bar{\nabla}_\mu q^\mu + (\rho + P)\Theta + 2q^\mu \dot{u}_\mu + \sigma_\nu^\mu \pi_\mu^\nu$ and $\gamma_\nu^\lambda \nabla_\mu T^{\mu\nu} = \gamma_\nu^\lambda \dot{q}^\nu + \bar{\nabla}^\lambda P + \bar{\nabla}_\alpha \pi^{\alpha\lambda} + \frac{4}{3}\Theta q^\lambda + (\sigma_\alpha^\lambda + \omega_\alpha^\lambda)q^\alpha + (\rho + P)\dot{u}^\lambda + \pi^{\alpha\lambda} \dot{u}_\alpha$. We get that

$$\dot{\rho} + \bar{\nabla}_\mu q^\mu + (\rho + P)\Theta + 2q^\mu \dot{u}_\mu + \sigma_\nu^\mu \pi_\mu^\nu = 0 \quad (1.87)$$

$$\gamma_\nu^\lambda \dot{q}^\nu + \bar{\nabla}^\lambda P + \bar{\nabla}_\alpha \pi^{\alpha\lambda} + \frac{4}{3}\Theta q^\lambda + (\sigma_\alpha^\lambda + \omega_\alpha^\lambda)q^\alpha + (\rho + P)\dot{u}^\lambda + \pi^{\alpha\lambda} \dot{u}_\alpha = 0. \quad (1.88)$$

These equations are valid for any energy-momentum tensor and any space-time geometry.

1.3.2.2 Perfect fluid

The perfect fluid is an especially interesting case for which $\pi_{\mu\nu} = 0$ and $q^\mu = 0$. The two previous equations (1.87) and (1.88) reduce, respectively, to the matter continuity equation and to the Euler equation

$$\dot{\rho} + (\rho + P)\Theta = 0, \quad (1.89)$$

$$(\rho + P)\dot{u}_\mu + \bar{\nabla}_\mu P = 0. \quad (1.90)$$

The term $\dot{u}_\mu = u^\alpha \nabla_\alpha u_\mu$ represents an acceleration that vanishes for a pressureless fluid ($P = 0$). Particles of such a pressureless fluid will hence follow geodesics.

For an irrotational flow, using (1.85) to express the Einstein tensor, (1.78) is of the form

$$(3) R = 2\kappa\rho - \frac{2}{3}\Theta^2 + 2\sigma^2 + 2\Lambda, \quad (1.91)$$

that we will call the *generalized Friedmann equation*.

Some conditions are often imposed on the pressure and the energy. The energy conditions introduced previously can be rewritten as

- null: $\rho \geq -P$,
- causal: $|\rho| \geq |P|$,
- null and causal: $|\rho| \geq |P|$ and $\rho < 0$ if and only if $\rho = -P$,
- weak: $\rho + P \geq 0$ and $\rho \geq 0$,
- strong: $\rho + 3P \geq 0$ and $\rho + P \geq 0$.

1.3.2.3 Electromagnetism

Electromagnetism is described by the antisymmetric tensor $F_{\mu\nu}$, called the Faraday tensor and satisfying

$$3F_{[\mu\nu;\alpha]} = F_{\mu\nu;\alpha} + F_{\nu\alpha;\mu} + F_{\alpha\mu;\nu} = 0. \quad (1.92)$$

$F_{\mu\nu}$ can be defined as the curl of the vector potential A_μ as

$$F_{\mu\nu} = A_{\nu,\mu} - A_{\mu,\nu} = A_{\nu,\mu} - A_{\mu,\nu}. \quad (1.93)$$

The energy-momentum tensor can be derived from the action

$$S_{em} = -\frac{1}{4} \int F_{\mu\nu} F^{\mu\nu} \sqrt{-g} d^4x, \quad (1.94)$$

which gives

$$T^{\mu\nu} = F^{\lambda\mu} F_\lambda^\nu - \frac{1}{4} g^{\mu\nu} F_{\alpha\beta} F^{\alpha\beta}. \quad (1.95)$$

The conservation equations then reduce to Maxwell's equations in the vacuum

$$\nabla_\mu F^{\nu\mu} = 0. \quad (1.96)$$

This equation can be rewritten in terms of the potential vector A^μ as

$$\nabla_\nu \nabla^\nu A_\mu - R_\mu^\nu A_\nu = 0, \quad (1.97)$$

where we chose the Lorentz gauge, defined by $\nabla_\mu A^\mu = 0$ and where the Ricci tensor arises from the commutation of the covariant derivatives. A source term may be added by introducing a coupling term $A^\mu j_\mu$ in the Lagrangian to obtain

$$\nabla_\mu F^{\nu\mu} = j^\nu. \quad (1.98)$$

An observer with tangent vector u^μ can define an electric and a magnetic field, respectively, by

$$E_\mu = F_{\mu\nu} u^\nu, \quad B_\mu = -\frac{1}{2} \epsilon_{\mu\nu\rho\sigma} F^{\rho\sigma} u^\nu. \quad (1.99)$$

These vectors satisfy $E^\mu u_\mu = B^\mu u_\mu = 0$ and we can express the electromagnetic field tensor as

$$F_{\mu\nu} = 2u_{[\mu} E_{\nu]} - \epsilon_{\mu\nu\rho\sigma} u^\rho B^\sigma. \quad (1.100)$$

Using this expression, we can decompose the energy-momentum tensor as the one for a fluid

$$T^{\mu\nu} = \frac{1}{2}(E^2 + B^2)u^\mu u^\nu + \frac{1}{6}(E^2 + B^2)\gamma^{\mu\nu} + 2u^{(\mu}\epsilon^{\nu)}{}^{\rho\sigma\lambda} u_\rho E_\sigma B_\lambda + \pi^{\mu\nu},$$

with $E^2 = E^\mu E_\mu$, $B^2 = B_\mu B^\mu$ and (recall $\pi_{\mu\nu}$ is traceless)

$$\pi^{\mu\nu} = \frac{1}{3}(E^2 + B^2)\gamma^{\mu\nu} - E^\mu E^\nu - B^\mu B^\nu.$$

We can then deduce that the energy density measured by a comoving observer is $\rho_{\text{em}} = \frac{1}{2}(E^2 + B^2)$ and that the equation of state for radiation is $P_{\text{em}} = \frac{1}{3}\rho_{\text{em}}$.

The trajectory of any particle of charge e and mass m in an electromagnetic field is of the form

$$u^\mu \nabla_\mu u^\nu = \frac{e}{m} E^\nu = \frac{e}{m} F^\nu{}_\mu u^\mu. \quad (1.101)$$

Finally, we can re-express Maxwell's equations, $\nabla_\mu F^{\nu\mu} = j^\nu$, in terms of E^μ and B^μ . Projecting along u_μ and along γ_μ^λ we get, respectively,

$$\bar{\nabla}_\mu E^\mu - 2\omega_\mu B^\mu = \epsilon, \quad (1.102)$$

$$\gamma_\mu^\lambda \dot{E}^\mu - \epsilon^{\alpha\lambda\rho\sigma} u_\alpha \bar{\nabla}_\rho B_\sigma = -\gamma_\nu^\lambda j^\nu - \frac{2}{3}\Theta E^\lambda + \sigma_\nu^\lambda E^\nu + \epsilon^{\alpha\lambda\rho\sigma} u_\alpha (\dot{u}_\rho B_\sigma + \omega_\rho E_\sigma), \quad (1.103)$$

with $\epsilon = -j^\mu u_\mu$. The same procedure applied to the equation $F_{[\mu\nu;\alpha]} = 0$ gives, respectively,

$$\bar{\nabla}_\mu B^\mu + 2\omega_\mu B^\mu = 0, \quad (1.104)$$

$$\gamma_\mu^\lambda \dot{B}^\mu + \epsilon^{\alpha\lambda\rho\sigma} u_\alpha \bar{\nabla}_\rho E_\sigma = -\frac{2}{3}\Theta B^\lambda + \sigma_\nu^\lambda B^\nu - \epsilon^{\alpha\lambda\rho\sigma} u_\alpha (\dot{u}_\rho E_\sigma - \omega_\rho B_\sigma). \quad (1.105)$$

Written in this form, these four equations give a relativistic generalization of the four Maxwell's equations, valid for any space-time geometry.

1.3.2.4 Scalar field

The action of a scalar field ϕ evolving in a potential $V(\phi)$ is given by

$$S_\phi = -\frac{1}{2} \int [\partial_\mu \phi \partial^\mu \phi + 2V(\phi)] \sqrt{-g} d^4x. \quad (1.106)$$

Varying this action with respect to the metric, we get the energy-momentum tensor as defined in (1.84)

$$T_{\mu\nu} = \partial_\mu \phi \partial_\nu \phi - g_{\mu\nu} \left[\frac{1}{2} \partial^\alpha \phi \partial_\alpha \phi + V(\phi) \right], \quad (1.107)$$

for which the conservation equation is simply the Klein–Gordon equation

$$\nabla_\mu \nabla^\mu \phi = V'(\phi). \quad (1.108)$$

We notice that the energy-momentum tensor of a scalar field can be decomposed similarly to that of a perfect fluid with velocity field

$$u^\mu = -\frac{1}{\psi} \nabla^\mu \phi, \quad \psi = \dot{\phi} = \sqrt{-\nabla^\mu \phi \nabla_\mu \phi},$$

(and $q^\mu = 0$, $\pi_{\mu\nu} = 0$) and with energy density and pressure

$$\rho_\phi = \frac{1}{2} \psi^2 + V(\phi), \quad P_\phi = \frac{1}{2} \psi^2 - V.$$

We can also note that $\Theta = -(\dot{\psi} + V')/\psi$, $\omega_{\mu\nu} = 0$, $a_\mu = -(\partial_\mu \psi + \dot{\psi} u_\mu)/\psi$ and $\sigma_{\mu\nu} = -\gamma_\mu^\lambda \gamma_\nu^\delta (\nabla_{(\lambda} \nabla_{\delta)} \phi)/\psi + \gamma_{\mu\nu} \nabla_\alpha [\psi^{-1} \nabla^\alpha \phi]/3$. In particular, it follows that the conservation equations reduce to

$$\dot{\rho} + \Theta(\rho + P) = 0 \iff \ddot{\phi} + \Theta \dot{\phi} + V' = 0,$$

for any Universe geometry.

1.3.3 ADM Hamiltonian formulation

To finish, we shall rewrite the gravitational action in the $1+3$ form provided by the Hamiltonian analysis by Arnowitt, Deser, and Misner [4].

We consider space-times that can be foliated by a continuous family of three-dimensional hypersurfaces, Σ_t . This means that there exists a smooth function \hat{t} on \mathcal{M} whose gradient never vanishes so that each hypersurface is a surface of constant \hat{t}

$$\Sigma_t \equiv \{p \in \mathcal{M}, \quad \hat{t}(p) = t\},$$

$\forall t \in \mathbb{R}$. Such space-times are called *globally hyperbolic* and they represent most space-times of astrophysical and cosmological interest. We assume that each hypersurface of the foliation is space-like and that it covers \mathcal{M} .

If we split the line-element as

$$ds^2 = -N^2 dt^2 + \gamma_{ij} (N^i dt + dx^i) (N^j dt + dx^j), \quad (1.109)$$

then the normal vector to Σ_t has components $n^\alpha = (1, -N^i)/N$, i.e. $n_\alpha = N(-1, 0, 0, 0)$. Calling t^α the vector defined by $t^\alpha \nabla_\alpha t = 1$, then

$$N = -t^\alpha n_\alpha, \quad N_\alpha = (g_{\alpha\beta} + n_\alpha n_\beta) t^\beta,$$

and $t^\alpha = N n^\alpha + N^\alpha$. This gives a geometrical interpretation to N and N_i : N is the *lapse function*, N^i the *shift vector* and γ_{ij} the spatial metric. It is thus clear that

the components of the metric are $g^{00} = -N^2 + \gamma_{ij} N^i N^j$, $g_{0i} = N_i$ and $g_{ij} = \gamma_{ij}$ and the components of the inverse metric are $g_{00} = -N^{-2}$, $g^{0i} = N^i/N^2$ and $g^{ij} = \gamma^{ij} - N^i N^j/N^2$. As t increases, we go from a hypersurface to another and we can view the four-dimensional space-time as the time evolution of a three-dimensional Riemannian space.

One can then deduce a simple relation between the determinants g of $g_{\mu\nu}$ and γ of γ_{ij} . Using Cramer's rule to compute the inverse metric, one gets that $g^{00} = \gamma/g$ so that

$$\sqrt{-g} = N \sqrt{\gamma}.$$

The Einstein-Hilbert action (1.80) can be rewritten as

$$S = \int \left[\int_{\Sigma_t} N \left({}^{(3)}R + K_{ij} K^{ij} - K^2 \right) \sqrt{\gamma} d^3x \right] dt.$$

This action has to be considered as a functional of the variables $q = (\gamma_{ij}, N, N^i)$ and their time derivatives. The extrinsic curvature is

$$K_{ij} = \frac{1}{2N} (\gamma_{ik} D_j N^k + \gamma_{jk} D_i N^k - \dot{\gamma}_{ij}),$$

so that the gravity sector derives from the Lagrangian density

$$\mathcal{L}[q, \dot{q}] = N \left({}^{(3)}R + K_{ij} K^{ij} - K^2 \right) \sqrt{\gamma}. \quad (1.110)$$

Since this Lagrangian does not depend on \dot{N} and \dot{N}^i , we conclude that the lapse function and the shift vector are not dynamical variables and will be associated to two constraint equations. The only dynamical variable is therefore γ_{ij} and its conjugate momentum is

$$\pi^{ij} \equiv \frac{\delta \mathcal{L}}{\delta \dot{\gamma}_{ij}} = \sqrt{\gamma} (K \gamma^{ij} - K^{ij}). \quad (1.111)$$

It follows that the Hamiltonian density is given by $\mathcal{H} = \pi^{ij} \dot{\gamma}_{ij} - \mathcal{L}$ and can be expressed as

$$\begin{aligned} \mathcal{H} = & -\sqrt{\gamma} \left[N \left({}^{(3)}R - K_{ij} K^{ij} + K^2 \right) + 2N^i (D_i K - D_j K_i^j) \right] \\ & + 2\sqrt{\gamma} D_j (K N^j - K_i^j N^i). \end{aligned}$$

The last term, being a divergence, does not contribute to the integral, so that the Hamiltonian reduces to

$$H = - \int_{\Sigma_t} (N C_0 - 2N^i C_i) \sqrt{\gamma} d^3x, \quad (1.112)$$

with

$$C_0 = {}^{(3)}R - K_{ij} K^{ij} + K^2 \quad C_i = -D_i K + D_j K_i^j. \quad (1.113)$$

It is a functional of $q = (\gamma_{ij}, N, N^i)$ and $p = \delta \mathcal{L}/\delta \dot{q} = (\pi^{ij}, 0, 0)$. The scalar curvature ${}^{(3)}R$ is a function of γ_{ij} and its spatial derivatives and the extrinsic curvature is a function of γ_{ij} and π^{ij} obtained from the inversion of (1.111)

$$K_{ij} = \frac{1}{\sqrt{\gamma}} \left(\frac{1}{2} \gamma_{kl} \pi^{kl} \gamma_{ij} - \gamma_{ik} \gamma_{jl} \pi^{kl} \right).$$

The Hamilton equations lead to two constraints, the ‘energy constraint’ $C_0 = 0$ and the ‘momentum constraint’ $C_i = 0$, and an evolution equation, which in the vacuum takes the form $\dot{\gamma}_{ij} = \delta H / \delta \pi^{ij} = -2NK_{ij} + D_i N_j + D_j N_i$. It is beyond the scope of this book to derive the Hamilton equation with matter. When matter is included, the constraint equations take the form

$$(3) R + K^2 - K_{ij} K^{ij} = 16\pi G\rho, \quad (1.114)$$

$$D_j K_i^j - D_i K = 8\pi G_N P_i, \quad (1.115)$$

with $\rho \equiv T_{\mu\nu} n^\mu n^\nu$ and $P_\beta = -(\delta_\beta^\mu + n_\beta n^\mu) T_{\mu\alpha} n^\alpha$. These relations can be obtained directly from the Einstein equations and the Gauss relation (1.76).

1.4 Geodesics in a curved space-time

1.4.1 Conserved quantities along a geodesic

To each symmetry one can associate a Killing vector, ξ^μ , under which the space-time metric remains invariant. Such a vector satisfies the equation deduced from (1.58), $\nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu = 0$. The existence of such vectors makes it possible to compute conserved quantities useful for the study of the solutions of Einstein’s equations.

Consider a geodesic with tangent vector u^μ . One can check that the quantity $u^\mu \xi_\mu$ is constant along the trajectory. Indeed,

$$u^\nu \nabla_\nu (u^\mu \xi_\mu) = \xi_\mu u^\nu \nabla_\nu u^\mu + u^\nu u^\mu \nabla_\nu \xi_\mu = 0.$$

The first term vanishes from the geodesics equation and the second term is the contraction of a symmetric tensor with an antisymmetric tensor.

Similarly, if $T^{\mu\nu}$ is an energy-momentum tensor then $T^{\mu\nu} \xi_\nu$ is a conserved vector

$$\nabla_\mu (T^{\mu\nu} \xi_\nu) = \xi_\nu \nabla_\mu T^{\mu\nu} + T^{\mu\nu} \nabla_\mu \xi_\nu = 0.$$

The first term vanishes from the conservation equation and the second is the contraction of a symmetric tensor with an antisymmetric tensor. This relation has an important implication. The vector $\hat{I}^\mu = T^{\mu\nu} \xi_\nu \sqrt{-g}$ satisfies by construction

$$\partial_\mu \hat{I}^\mu = \sqrt{-g} \nabla_\mu (T^{\mu\nu} \xi_\nu) = 0.$$

Integrating over a volume V , we can hence conclude that

$$\int_V \nabla_\mu T^{\mu\nu} \xi_\nu \sqrt{-g} d^3x = \int_V \partial_\mu \hat{I}^\mu d^3x = 0.$$

The quantity \hat{Q}_0 , defined by

$$\hat{Q}_0 = \int_V \hat{I}^0 d^3x,$$

follows the conservation equation

$$\partial_0 \hat{Q}^0 = - \int_{\partial V} \hat{I}' dS_i.$$

It follows that $\partial_0 \hat{Q}^0 = 0$ if the boundary term vanishes.

1.4.2 Geodesic deviation equation

We consider a family $x^\mu(\lambda, s)$ of neighbouring geodesics, where λ is the affine parameter along the geodesics and s is a label. The vector

$$n^\mu = \frac{\partial x^\mu}{\partial s}$$

measures the separation between two neighbouring geodesics. It can be chosen in the plane orthogonal to the geodesic flow, i.e. such that $u^\mu n_\mu = 0$ (otherwise one can replace it by $\gamma_\nu^\mu n^\nu$). One can check that

$$n^\mu \nabla_\mu u^\nu = n^\mu \partial_\mu u^\nu + \Gamma_{\mu\alpha}^\nu u^\alpha n^\mu = \frac{\partial^2 x^\nu}{\partial \lambda \partial s} + \Gamma_{\mu\alpha}^\nu u^\alpha n^\mu = u^\mu \nabla_\mu n^\nu.$$

The relative acceleration between two neighbouring geodesics is defined by

$$a^\mu = \frac{d^2 n^\mu}{ds^2} = u^\alpha \nabla_\alpha (u^\beta \nabla_\beta n^\mu). \quad (1.116)$$

Using the previous property we obtain

$$a^\mu = u^\alpha \nabla_\alpha (n^\beta \nabla_\beta u^\mu) = (u^\alpha \nabla_\alpha n^\beta) \nabla_\beta u^\mu + u^\alpha n^\beta \nabla_\alpha \nabla_\beta u^\mu. \quad (1.117)$$

To evaluate the first term, note that $n^\mu \nabla_\mu (u^\nu \nabla_\nu u^\alpha) = 0$ since u^μ satisfies the geodesics equation. This means that $(n^\mu \nabla_\mu u^\nu) \nabla_\nu u^\alpha + n^\mu u^\nu \nabla_\mu \nabla_\nu u^\alpha = 0$ that can be rewritten, switching n^μ and u^μ in the first term and switching the dummy indices μ and ν in the second one, as $(u^\mu \nabla_\mu n^\nu) \nabla_\nu u^\alpha + n^\nu u^\mu \nabla_\nu \nabla_\mu u^\alpha = 0$. Using this result to evaluate the second term of (1.117), we obtain, after commuting the two partial derivatives

$$a^\mu = R^\mu_{\nu\alpha\beta} u^\nu u^\alpha n^\beta = -R^\mu_{\nu\beta\alpha} u^\nu n^\beta u^\alpha. \quad (1.118)$$

This equation confirms the heuristic argument on the relation between curvature and deviation of nearby geodesics. Even if we work in a free-falling referential (where gravity locally vanishes) the Riemann tensor does not vanish (because the derivatives of the Christoffel symbols a priori do not cancel) and one cannot avoid the focusing or defocusing of neighbouring geodesics.

1.4.2.1 Covariant approach

We decompose the vector n^μ as $n^\mu = \ell e^\mu$ with $e^\mu e_\mu = 1$ (thus, $\dot{e}^\mu e_\mu = 0$) and $e^\mu u_\mu = 0$. We can then express $\dot{n}^\mu = u^\nu \nabla_\nu n^\mu$ either as $\dot{n}^\mu = \ell \dot{e}^\nu + \dot{\ell} e^\nu$ or as $\dot{n}^\mu = n^\nu \nabla_\nu u^\mu$. By comparing these two expressions and projecting on e^μ , we get

$$\frac{\dot{e}}{e} = \frac{1}{3}\Theta + \sigma_{\mu\nu}e^\mu e^\nu, \quad (1.119)$$

that we call the generalized Hubble law. As for the orthogonal projection, it gives

$$\gamma_\lambda^\mu \dot{e}^\lambda = [\sigma_\lambda^\mu - (\sigma_{\alpha\beta}e^\alpha e^\beta)\gamma_\lambda^\mu - \omega_\lambda^\mu] e^\lambda, \quad (1.120)$$

which describes the change in the direction of observation of a distant galaxy, for instance.

1.4.2.2 Optics and redshift

We now consider an electromagnetic wave with a potential vector of the form

$$A_\mu = C_\mu e^{i\varphi}.$$

The Lorentz gauge condition imposes

$$C^\mu \partial_\mu \varphi = 0.$$

If we assume that the amplitude C_μ varies slowly compared to the phase φ and that the wavelength is small compared to the space-time curvature length scale, we can neglect the curvature term in (1.97) so that Maxwell's equations impose

$$\nabla^\mu \nabla_\mu \varphi = 0, \quad \partial_\mu \varphi \partial^\mu \varphi = 0.$$

For any function φ , the vector $k^\mu = \partial^\mu \varphi$ is orthogonal to the wave surface on which $\varphi = \text{const.}$; k^μ is the wavevector. If we differentiate the second equation and use the fact that $\nabla_\nu \partial_\mu \varphi = \nabla_\mu \partial_\nu \varphi$, we get that light is a transverse wave propagating along null geodesics

$$\begin{aligned} A_\mu &= C_\mu e^{i\varphi}, & k^\mu &= \partial^\mu \varphi, \\ C_\mu k^\mu &= 0, & k_\mu k^\mu &= 0, & k^\mu \nabla_\mu k^\nu &= 0. \end{aligned} \quad (1.121)$$

In a Minkowski space-time, $k^\mu = (-E, \mathbf{k})$, where E is the photon energy and \mathbf{k} its momentum. An observer that follows a trajectory with tangent vector u^μ will measure an energy (or equivalently, a frequency) given by

$$E = \hbar\omega = -k^\mu u_\mu.$$

A photon emitted with a frequency ω_{em} will be received with a frequency ω_{rec} (see Fig. 1.8). These two frequencies will be related by

$$\frac{\lambda_{\text{rec}}}{\lambda_{\text{em}}} = \frac{\omega_{\text{em}}}{\omega_{\text{rec}}} = \frac{(k_\mu u^\mu)_{\text{em}}}{(k_\mu u^\mu)_{\text{rec}}} \equiv 1 + z. \quad (1.122)$$

This relation will be very important in cosmology in order to define the notion of redshift, z . The wavevector k^μ can always be decomposed as $k^\mu = E(u^\mu + e^\mu) = dx^\mu/dp$, with $u^\mu e_\mu = 0$ and $e^\mu e_\mu = 1$ and where p is the affine parameter along the photon trajectory. It follows that $dE/dp = -k^\mu \nabla_\mu (k_\nu u^\nu)$ and hence

$$\frac{1}{\lambda} \frac{d\lambda}{dp} = -\frac{1}{E} \frac{dE}{dp} = \left(\frac{1}{3}\Theta + \dot{u}_\mu e^\mu + \sigma_{\mu\nu}e^\mu e^\nu \right) E. \quad (1.123)$$

This expression points out the anisotropic contributions to the redshift coming from the shear. This expression is valid for any space-time.

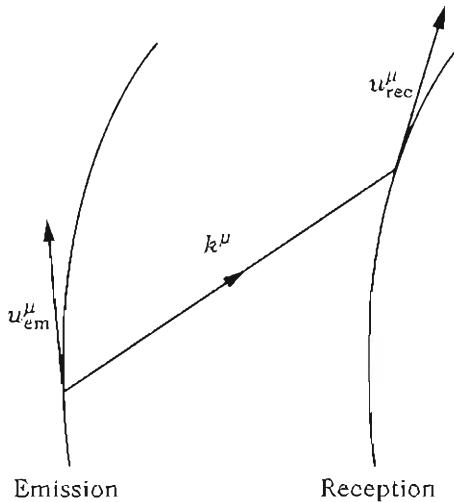


Fig. 1.8 A photon emitted by a comoving observer with 4-velocity u_{em}^μ , propagates along a null geodesic and is received by a comoving observer with 4-velocity u_{rec}^μ .

1.5 Weak-field regime

1.5.1 Newtonian limit

An undetermined parameter appears in the Einstein equations, the constant κ . To fix it, we consider the weak-field limit of the Einstein equations in order to understand how Newtonian gravity is recovered. We expand the space-time metric around a Minkowski space-time, in Minkowski coordinates, as

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \quad (1.124)$$

where $h_{\mu\nu}$ is a small perturbation.

1.5.1.1 Linearized Einstein equations

From the previous decomposition, we can conclude that to first order in $h_{\mu\nu}$,

$$g^{\mu\nu} = \eta^{\mu\nu} - h^{\mu\nu},$$

and that $h^{\mu\nu} = \eta^{\alpha\mu}\eta^{\beta\nu}h_{\alpha\beta}$ and we set $h = h_\alpha^\alpha$. To linear order in $h_{\mu\nu}$, the Christoffel symbols are given by

$$\Gamma_{\mu\nu}^\alpha = \frac{1}{2}\eta^{\alpha\lambda} [2h_{\lambda(\mu,\nu)} - h_{\mu\nu,\lambda}].$$

Accordingly, the Ricci tensor is of the form $R_{\mu\nu} = \partial_\alpha\Gamma_{\mu\nu}^\alpha - \partial_\mu\Gamma_{\alpha\nu}^\alpha$, so that

$$R_{\mu\nu} = \frac{1}{2} [2\partial^\lambda\partial_{(\mu}h_{\nu)\lambda} - \square h_{\mu\nu} - \partial_{\mu\nu}h], \quad (1.125)$$

where $\square = \eta^{\alpha\beta}\partial_\alpha\partial_\beta$ is the d'Alembertian in Minkowski space. Contracting the previous expression with $\eta^{\mu\nu}$, we get the scalar curvature

$$R = h_{,\lambda\sigma}^{\lambda\sigma} - \square h.$$

The Einstein tensor is given by $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}R\eta_{\mu\nu}$ so that the Einstein equations read

$$2\partial^\lambda\partial_{(\mu}\bar{h}_{\nu)\lambda} - \square\bar{h}_{\mu\nu} - \eta_{\mu\nu}\partial_\alpha\partial_\beta\bar{h}^{\alpha\beta} = 2\kappa T_{\mu\nu} \quad (1.126)$$

with

$$\bar{h}_{\mu\nu} = h_{\mu\nu} - \frac{1}{2}h\eta_{\mu\nu}.$$

In particular, we can check that $\bar{h} = -h$ and that $h_{\mu\nu} = \bar{h}_{\mu\nu} - \frac{1}{2}\bar{h}\eta_{\mu\nu}$.

1.5.1.2 Gauge freedom

In general relativity, there exist a gauge freedom related to the arbitrary choice of coordinate systems: two different metrics can describe the same space-time. In the linearized limit, two perturbations can represent the same state if they are related by a change of coordinates. After a change of coordinates of the form $x^\mu \rightarrow x^\mu + \xi^\mu$, the metric perturbation transforms as

$$h_{\mu\nu} \rightarrow h_{\mu\nu} - \mathcal{L}_\xi\eta_{\mu\nu}, \quad \bar{h}_{\mu\nu} \rightarrow \bar{h}_{\mu\nu} - \mathcal{L}_\xi\eta_{\mu\nu} + \eta_{\mu\nu}\partial_\alpha\xi^\alpha,$$

where the Lie derivative is given by $\mathcal{L}_\xi\eta_{\mu\nu} = 2\partial_{(\mu}\xi_{\nu)}$.

Using these transformations, the linearized Einstein equation (1.126), takes the remarkably simple form

$$\boxed{\square\bar{h}_{\mu\nu} = -2\kappa T_{\mu\nu}} \quad (1.127)$$

if one chooses ξ^μ such that $\square\xi^\mu = \partial_\nu\bar{h}^{\mu\nu}$. This means that after the change of variable, the metric perturbation satisfies the gauge condition

$$\partial_\nu\bar{h}^{\mu\nu} = 0. \quad (1.128)$$

This gauge is known as the *de Donder gauge* or *harmonic gauge*. It is analogous to the Lorentz gauge in electromagnetism.

1.5.1.3 Solutions of Einstein's equations

In the harmonic gauge, the solution of Einstein's equations can be found using the retarded potential method, just like in electromagnetism,

$$\bar{h}_{\mu\nu}(x, t) = \frac{\kappa}{2\pi} \int \frac{T_{\mu\nu}(x', t - |x - x'|/c)}{|x - x'|} d^3x'.$$

The general solution is therefore of the form $\bar{h}_{\mu\nu}^{(\text{gen})} = \bar{h}_{\mu\nu} + \bar{h}_{\mu\nu}^{(H)}$, where $\bar{h}_{\mu\nu}^{(H)}$ is a solution of the homogeneous equations $\square\bar{h}_{\mu\nu}^{(H)} = 0$ with $\partial^\nu\bar{h}_{\mu\nu}^{(H)} = 0$.

1.5.1.4 Geodesic equation

For a classical fluid, and to lowest order in the velocity, only the component T_{00} gives a contribution (since $P \ll pc^2$). T_{0i} and T_{ij} should be taken into consideration only at the first post-Newtonian (1PN) order, i.e. at order v^2/c^2 . The form of the general solution of the linearized Einstein equations (1.127) implies that the only non-vanishing terms are \bar{h}_{00} and \bar{h}_{0i} . They are given by

$$\bar{h}_{00} = \frac{\kappa c^2}{2\pi} \int \frac{\rho}{|x - x'|} d^3x', \quad \bar{h}_{0i} = \frac{\kappa c}{2\pi} \int \frac{\rho u_i}{|x - x'|} d^3x'.$$

It follows that all diagonal components of $h_{\mu\nu}$ are equal to $\bar{h}_{00}/2$.

We now consider the trajectory of a particle with velocity small compared to the speed of light ($v/c \ll 1$). As we have seen before, to lowest order in h the tangent vector to the worldline is given by

$$u^\mu = \frac{dx^\mu}{d\tau} = \gamma^{-1} \frac{dx^\mu}{dt},$$

with $v^i \equiv dx^i/dt$ and $\gamma = dt/d\tau = \sqrt{1 - v^2/c^2}$, so that we have $u^\mu = \gamma^{-1}(c, v^i)$. Using the relation $d^2x^\mu/dt^2 = d(\gamma u^\mu)/dt = u^\mu d\gamma/dt + \gamma du^\mu/dt$, the spatial component of the geodesic equation becomes

$$\frac{d^2x^i}{dt^2} = -\Gamma_{\alpha\beta}^i \frac{dx^\alpha}{dt} \frac{dx^\beta}{dt} + \Gamma_{\alpha\beta}^0 \frac{dx^\alpha}{dt} \frac{dx^\beta}{dt} \frac{dx^i}{dt}.$$

To linear order in h and v , this equation is of the form

$$\frac{d^2x^i}{dt^2} = -c^2 \left[\Gamma_{00}^i + \left(\Gamma_{00}^0 \delta_j^i - \frac{2}{c} \Gamma_{0j}^i \right) v^j \right]. \quad (1.129)$$

1.5.1.5 Determination of κ

Let us work in a static gravitational field. The perturbation $h_{\mu\nu}$ only depends on the position x^i . Equation (1.129) then simplifies to

$$\frac{d^2x^i}{dt^2} = -\frac{c^2}{2} \delta^{ij} \partial_j h_{00}.$$

The first term represents the coordinate acceleration. In the second term, we recognize the gradient of a scalar that we identify with the gravitational potential $2\phi = c^2 h_{00}$, so that we recover the Newtonian form

$$\frac{d^2x^i}{dt^2} = -\delta^{ij} \partial_j \phi, \quad h_{00} = 2 \frac{\phi}{c^2}.$$

We can therefore go back to the Einstein equations and compare them in this limit with the Poisson equation that determines the gravitational potential in Newtonian mechanics. This allows one to identify the value of κ such that the weak-field limit of relativity is consistent with Newtonian gravity. From the previous equation, we deduce that $\bar{h}_{00} = 4\phi/c^2$. The energy-momentum tensor has for main component $T_{00} = \rho c^2$

and the limit of the Einstein equation is therefore $\Delta\phi = \kappa pc^4/2$. Comparing this equation with the Poisson equation $\Delta\phi = 4\pi G_N\rho$, it follows that

$$\boxed{\kappa = \frac{8\pi G_N}{c^4}}, \quad (1.130)$$

where G_N is the Newton constant.

1.5.1.6 Analogy with electromagnetism

In the previous expressions, the linearized equations of general relativity are very similar to the equations of electromagnetism. One can take the analogy a step further by defining

$$A_i = c \bar{h}_{0i}.$$

The geodesic equation can then be rewritten as

$$\ddot{x} = -\nabla\phi - \dot{A} + v \wedge (\nabla \wedge A) + 3\dot{\phi}v.$$

In a static space-time, for which $\dot{\phi} = 0$, the previous expression reduces to a form similar to the one in electromagnetism. The Einstein equations are hence similar to the Maxwell equations if we introduce

$$\begin{aligned} g &= -\nabla\phi - \frac{1}{c}\partial_t A, & B &= \nabla \wedge A, \\ \nabla \cdot B &= 0, & \nabla \wedge B &= 16\pi G_N \rho v + \frac{1}{c}\partial_t E, \\ \nabla \wedge g &= -\frac{1}{c}\partial_t B, & \nabla \cdot g &= 4\pi G_N \rho. \end{aligned} \quad (1.131)$$

The following list summarizes some aspects of the analogy, which cannot unfortunately be pushed much further.

| | electromagnetism | linearized relativity |
|---|---|--|
| Variables | A_μ | $h_{\mu\nu}$ |
| Gauge transformations | $A_\mu \rightarrow A_\mu + \partial_\mu \chi$ | $\bar{h}_{\mu\nu} \rightarrow \bar{h}_{\mu\nu} - \mathcal{L}_\xi \eta_{\mu\nu} + \eta_{\mu\nu} \partial_\alpha \xi^\alpha$, |
| Gauge condition | $\partial_\mu A^\mu = 0$ | $\partial_\mu \bar{h}^{\mu\nu} = 0$ |
| Field equations | $\square A_\mu = -j_\mu$ | $\square \bar{h}_{\mu\nu} = -2\kappa T_{\mu\nu}$ |
| Restrictions on the other gauge transformations | $\square \chi = 0$ | $\square \xi^\mu = 0$ |

1.5.1.7 SVT decomposition

To go further in the study of perturbations, let us decompose $h_{\mu\nu}$ as

$$h_{00} = -2A, \quad h_{0i} = \partial_i B + \bar{B}_i, \quad h_{ij} = 2C\delta_{ij} + 2\partial_{ij}E + 2\partial_{(i}\bar{E}_{j)} + \bar{E}_{ij}$$

with

$$\partial_i \bar{B}^i = 0, \quad \partial_i \bar{E}^i = 0, \quad \partial_i \bar{E}^{ij} = 0, \quad \bar{E}_i^i = 0.$$

The 10 components of the tensor $h_{\mu\nu}$ are decomposed into 4 scalar perturbations (A, B, C, E), 2 vectors perturbations (\bar{B}^i, \bar{E}^i) that have $6 - 2 = 4$ independent components and 1 tensor perturbation \bar{E}_{ij} that has $6 - 4 = 2$ independent components.

The important property of this decomposition, which will be used in the cosmological context in Chapter 5, is that these three types of perturbations decouple (see Ref. [11]).

As previously, we should take into account the effect of an arbitrary change of coordinates. In order to do so, we decompose the displacement vector field ξ^μ as

$$\xi^0 = -T, \quad \xi^i = -\partial^i L - \bar{L}^i.$$

Under such coordinate transformation, the metric transforms as $g_{\mu\nu} \rightarrow g_{\mu\nu} - \mathcal{L}_\xi g_{\mu\nu}$ [see (1.58)]. It follows that⁵

$$A \rightarrow A + \dot{T}, \quad B \rightarrow B + \dot{L} - T, \quad C \rightarrow C, \quad E \rightarrow E + L$$

for the scalar modes,

$$\bar{B}_i \rightarrow \bar{B}_i + \dot{\bar{L}}_i, \quad \bar{E}_i \rightarrow \bar{E}_i + \bar{L}_i,$$

for the vector modes, and the tensor modes remain unchanged

$$\bar{E}_{ij} \rightarrow \bar{E}_{ij}.$$

The tensor modes are therefore gauge invariant since they do not depend on the choice of the coordinate system. This is not the case for the vector and scalar modes. However, we can define a combination of these modes that are gauge invariant. For the scalar modes, we define

$$\Phi = A + \dot{B} - \ddot{E}, \quad \Psi = -C,$$

and for the vector modes

$$\bar{\Phi}_i = \dot{\bar{E}}_i - \bar{B}_i.$$

We therefore have defined 2 scalar quantities and 1 vector quantity (2 degrees of freedom) which are gauge invariant. All together, we therefore have $6 = 10 - 4$ degrees of freedom, once the 4 arbitrary degrees of freedom related to the gauge choice have been absorbed.

1.5.2 Gravitational waves in an empty space-time

1.5.2.1 Propagation equations

For the scalar modes, the Einstein equations $G_{\mu\nu} = 0$, imply that $\Delta\Psi = 0$ (00 component), $\partial_i \dot{\Psi} = 0$ (0i component) and $\Phi - \Psi = 0$ (ij component). The only regular solution of the two first equations is $\Psi = 0$ and therefore

$$\Phi = \Psi = 0,$$

which means that no scalar mode can propagate.

For the vector modes, the component 0i implies that $\Delta\bar{\Phi}_i = 0$, the only regular solution of which is $\bar{\Phi}_i = 0$. Just as for scalar modes, no vector modes can propagate.

⁵To first order in ξ , $h_{\mu\nu} \rightarrow h_{\mu\nu} - \mathcal{L}_\xi \eta_{\mu\nu}$ and we just need to evaluate $\mathcal{L}_\xi \eta_{00} = 2\dot{T}$, $\mathcal{L}_\xi \eta_{0i} = \partial_i(T - \dot{L}) - \dot{\bar{L}}^i$ and $\mathcal{L}_\xi \eta_{ij} = -2\Delta L \delta_{ij} - 2\partial_{(i}\bar{L}_{j)}$.

The situation is different for tensor modes. The ij component of the Einstein equations implies that $\square \bar{E}_{ij} = 0$.

The only perturbations that can propagate in a Minkowski space-time are the gravitational waves and they satisfy

$$\square \bar{E}_{ij} = 0, \quad \bar{E}_i^i = 0, \quad \partial_i \bar{E}^{ij} = 0. \quad (1.132)$$

The three conditions $\Phi = \Psi = \bar{\Phi}_i = 0$ define a gauge equivalence class. We can choose a gauge in this family by imposing some conditions on the perturbations A, B, C, E and \bar{B}_i, \bar{E}_i . For instance, setting $E = B = 0$ and $\bar{E}_i = 0$, we define what is called a transverse and traceless (TT) gauge in which the metric is completely determined. In this case, the only non-vanishing component of $h_{\mu\nu}^{(TT)}$ is \bar{E}_{ij} .

1.5.2.2 Plane waves

Let us consider a plane wave propagating along the axis Oz in the TT gauge. Since the wave is transverse and traceless, it satisfies $h_{zz}^{(TT)} = h_{zx}^{(TT)} = h_{zy}^{(TT)} = 0$. Therefore, it has only two degrees of freedom (two polarizations) that should satisfy $h_{xx}^{(TT)} = -h_{yy}^{(TT)}$ and $h_{xy}^{(TT)} = h_{yx}^{(TT)}$. We hence can decompose it as follows

$$h_{ij}^{(TT)} = E_+(ct - z)\varepsilon_{ij}^+ + E_\times(ct - z)\varepsilon_{ij}^\times,$$

with the two polarization tensors being defined by

$$\varepsilon_{ij}^+ = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \varepsilon_{ij}^\times = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The solution of the propagation equation is $E_{+/\times} = A_{+/\times} \cos[k(z - ct) + \varphi]$, where φ is a phase.

1.5.2.3 Motion of a test particle

Consider the geodesic equation of a test particle in the gravitational field of a gravitational wave in the TT gauge. Since to leading order in perturbation we have $\Gamma_{00}^i = 0$, a particle initially at rest will remain at rest.

Of course, this does not mean that nothing happens, but rather that the frame of reference is comoving with the test particle. To see if anything happens, we should look at the relative motion of two neighbouring particles, which can be done using the geodesic deviation equation (1.118). The relative acceleration is given by $a^\mu = -R^\mu_{\nu\rho\sigma} u^\nu n^\rho u^\sigma$. To leading order in v/c , we have $u^\mu = (1, 0)$ and $n^\mu = (0, n^i)$, $R_{a0j0} = -\partial_t^2 h_{ij}^{(TT)}/2$ so that

$$\frac{d^2 n^i}{dt^2} = \frac{1}{2} \partial_t^2 h_j^{(TT)i} n^j.$$

Substituting the expression for $h_{ij}^{(TT)}$ from the previous section, we can easily integrate this expression. Figure 1.9 illustrates the deformation of a ring of particles under

the influence of a gravitational plane wave propagating along the axis z for each polarization case.

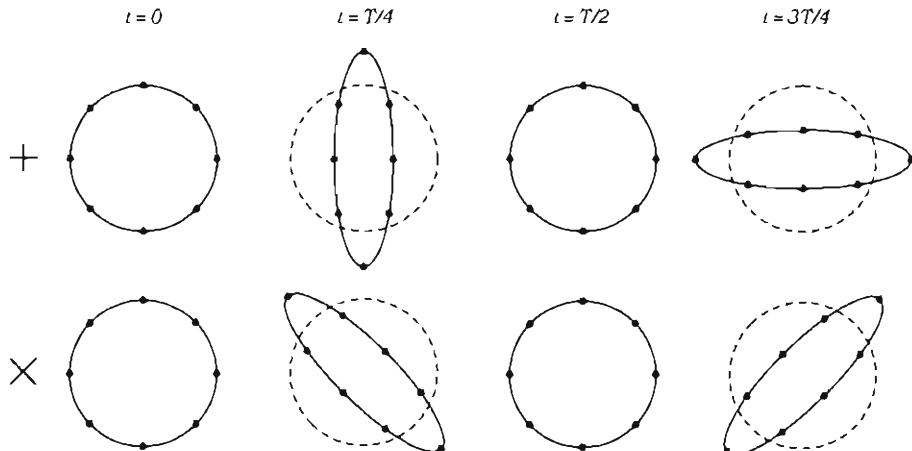


Fig. 1.9 Effects of a gravitational plane wave propagating along the axis z on a ring of particles located in the plane xy , depending on the wave polarization. The four snapshots are parameterized by the period T corresponding to a gravitational wave with wavelength k^{-1} .

1.6 Tests of general relativity

The cosmological models we will construct will strongly rely on the application of general relativity to the Universe as a whole. It is hence important to understand the limits of validity of our theory of gravitation.

We summarize here the principles of the main tests of gravitation; all constraints are summarized in the table at the end of this section. This question is presented in detail in Refs. [8, 12, 13].

1.6.1 Test of the equivalence principle

1.6.1.1 Universality of free fall

The mass of any body comes from different contributions (rest energy, electromagnetic or nuclear binding energy, etc.). If one of these energies contributed differently to the inertial mass than to the gravitational mass, we would observe a violation of the universality of free fall. To quantify this possible difference, we write

$$m_g = m_i + \sum_i \frac{1}{c^2} \eta^i E^i, \quad (1.133)$$

where for each interaction i , E^i is the associated binding energy, and η^i is a small dimensionless parameter that quantifies the amplitude of the violation of the universality of free fall associated to this interaction. The acceleration of any test body in an external gravitational field is given by

$$\mathbf{a} = \left(1 + \sum_i \eta^i \frac{E^i}{m_i c^2} \right) \mathbf{g},$$

so that measuring the difference between the acceleration of two different bodies allows us to define the *Eötvös parameter* η by

$$\eta = 2 \frac{|\mathbf{a}_1 - \mathbf{a}_2|}{|\mathbf{a}_1 + \mathbf{a}_2|} = \sum_i \eta^i \left(\frac{E^i}{m_i c^2} \Big|_1 - \frac{E^i}{m_i c^2} \Big|_2 \right), \quad (1.134)$$

to leading order in η^i . Two methods have mainly been used to constrain the value of the parameter η :

1. The first method uses the period T of a pendulum of length L ,

$$T = 2\pi \sqrt{\frac{m_1}{m_c}} \sqrt{\frac{L}{g}},$$

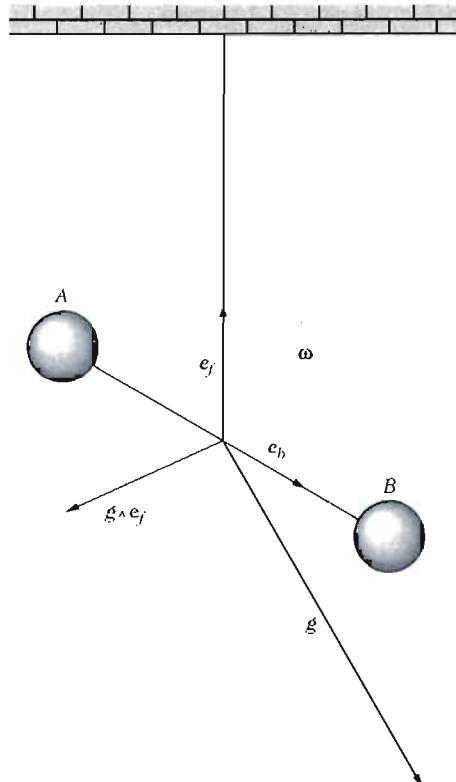


Fig. 1.10 Principle of Eötvös balance. \mathbf{g} is the external gravitational field and $\boldsymbol{\omega}$ is the angular velocity of the measurement device. The unit vectors \mathbf{e}_f and \mathbf{e}_b are, respectively, parallel to the torsion fibre and to AB .

so that

$$\eta = 2 \frac{|T_1 - T_2|}{T_1 + T_2}.$$

2. The second method uses a torsion balance (see Fig. 1.10). Two objects of different compositions, A and B, are attached at the end of a horizontal beam of length $2L$ suspended on a torsion fibre. The balance has an angular velocity ω , coming, for instance, from the Earth rotation; the force on each mass is therefore

$$\mathbf{F} = m_A \mathbf{g} + m_B \mathbf{c},$$

where c is the centrifugal acceleration. We can conclude from this expression that the wire tension is $\mathbf{T} = \mathbf{F}_A + \mathbf{F}_B$ and that the torque applied on the torsion fibre is $\mathbf{M} = L \mathbf{e}_b \wedge (\mathbf{F}_B - \mathbf{F}_A)$. There is therefore a torque M directed along the same direction as the tension. Performing the scalar product $\mathbf{M} \cdot \mathbf{T}$, we get that

$$M = \eta \frac{L}{|\mathbf{g} + \mathbf{c}|} (\mathbf{e}_b \wedge \mathbf{c}) \cdot \mathbf{g}.$$

This device was invented by the baron Roland von Eötvös (see Ref. [14]). In the original experiment, \mathbf{g} corresponded to the Earth's gravity and the rotation was that of the Earth.

The pendulum and torsion balance can test the weak equivalence principle because the gravitational binding energy of the test objects are negligible compared to the other energies. In order to test the strong equivalence principle, we should consider objects with an important gravitational binding energy such as celestial objects. This can be done by comparing the relative motion of the Earth and the Moon within the gravitational field of the Sun. There is therefore an extra acceleration with expression

$$\delta a = \frac{G_N M_{\odot}}{R^2} \eta \left(\left. \frac{E^{(\text{grav})}}{mc^2} \right|_{\text{Earth}} - \left. \frac{E^{(\text{grav})}}{mc^2} \right|_{\text{Moon}} \right) e_T,$$

where e_T is a unit vector directed from the Sun to the Earth and R the radius of the Earth orbit. This implies that the orbit of the Moon around the Earth is polarized in the direction of the Sun, with maximal amplitude

$$\delta r_{\max} = - \frac{3\delta a}{2(\omega_{\text{Earth}} \omega_{\text{Moon}})}.$$

Neglecting the gravitational energy of the Moon compared to that of the Earth, which is considered to be a homogeneous ball of radius R and mass m , it follows that $E^{(\text{grav})}/mc^2 = -3Gm/5Rc^2$. A more precise computation gives $E^{(\text{grav})}/mc^2 = -4.6 \times 10^{-10}$ for the Earth and -0.2×10^{-10} for the Moon, from which we conclude that

$$\delta a \leq 2.6 \times 10^{-12} \eta \text{ m} \cdot \text{s}^{-2}, \quad \delta r_{\max} \leq -14 \eta \text{ m}.$$

These measurements have been made possible, mainly owing to the Lunar Laser Ranging (LLR) experiment that measured the position of the Moon relative to the Earth with a precision of 1 cm over about thirty years.

1.6.1.2 Local Lorentz invariance

There exist few clean tests of the local Lorentz invariance. In particular, many of the high-energy tests cannot usually distinguish the specific effects of Lorentz invariance breaking from the complications due to weak and strong interactions.

The best constraint is the one obtained from the Hughes–Drever experiment [15]. The experiment observes the ground state of the lithium-7, with quantum number $J = 3/2$. In the presence of a magnetic field, this state splits into four equally spaced levels. Any perturbation associated with the existence of a preferred direction, would modify the spacing between the levels. This sets a limit on the potential anisotropy of the inertial mass δm_l^{ij} of the lithium nucleus that can be parameterized as

$$\delta m_l^{ij} = \sum_A \frac{1}{c^2} \delta_A E_A, \quad (1.135)$$

where δ_A is a dimensionless parameter that characterizes the amplitude of the anisotropy induced from the interaction A with binding energy E_A . In the experiment with the lithium-7, we get $|\delta m_l^{ij} c^2| \lesssim 1.7 \times 10^{-16}$ eV.

1.6.1.3 Local position invariance

The first test of local position invariance is related to the Einstein effect, i.e. to the fact that clocks run more slowly in the presence of a gravitational field. Consider an atomic transition at point A emitting a photon towards an observer located in B . According to the previous sections, the photon is received with a redshift $(1+z) = (u^\mu k_\mu)_B / (u^\mu k_\mu)_A$, implying that

$$z = \frac{\Delta\phi}{c^2},$$

where ϕ is the gravitational potential. If there is a space dependence, this relation should not be valid any longer. We would therefore expect

$$z = (1 + \alpha) \frac{\Delta\phi}{c^2}, \quad (1.136)$$

where α is a parameter that vanishes in general relativity. Several experiments have been performed to measure the redshift, either by sending atomic clocks in orbit or by observing solar spectra.

Another method consists of testing the variation of the fundamental constants. Numerous tests have been performed mainly on the fine structure constant, on the electron to proton mass ratio, and on the gravitational constant. These experiments go from local scales to the early Universe. We will come back to these tests in later chapters (see also the review Ref. [16]).

1.6.1.4 Fifth force

To finish, we point out tests on the potential presence of a fifth force. In these experiments, the gravitational potential is supposed to be modified by a Yukawa potential of range λ

$$U = \frac{G_\infty m_1 m_2}{r} \left(1 + \alpha_{12} e^{-r/\lambda} \right), \quad (1.137)$$

where α_{12} characterizes the amplitude of the deviation. The constraints on the pair (α_{12}, λ) are summarized in Fig. 1.11. In the laboratory, experiments using Cavendish balances can test Newton's law up to a hundred micrometers [17]. The constraints go up to scales of the order of the Solar System size by studying the motion of celestial objects.

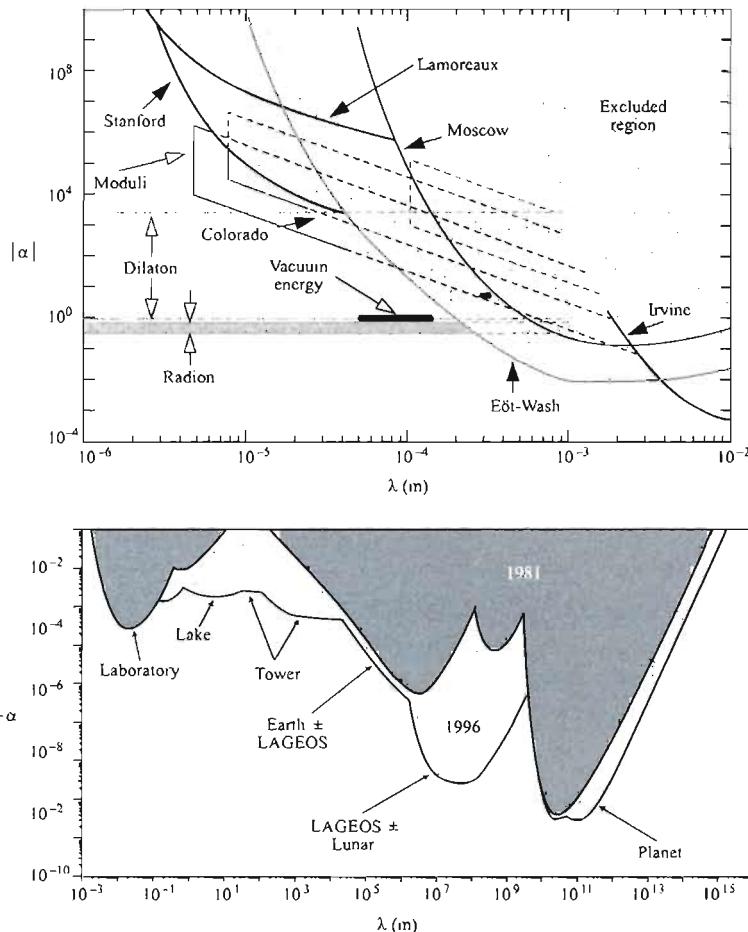


Fig. 1.11 Constraints on the existence of a fifth force deriving from a Yukawa potential. The parameter α represents the parameter α_{12} of (1.137), which we assume to be the same for all bodies. The bottom graph represents a zoom on the smallest tested scales. The diagram presents the predicted deviations expected from some models, such as the presence of extra dimensions, axions or dilaton, etc... These different models will be presented in later chapters of this book (quote from Ref. [18]).

1.6.2 Tests in the Solar System

We will now concentrate on the tests of the geometrical structure of space-time and of Einstein's equations. These tests rely mainly on the study of geodesics in the Solar System. We will focus on the case of Schwarzschild space-time that describes the metric generated by a massive body. We may, however, point out that many modern tests (such as LLR, LAGEOS or GPB experiments), can only be described with a N -body metric, i.e. beyond the Schwarzschild metric, which is beyond the scope of this book (see Refs. [1–4] for further details).

1.6.2.1 Schwarzschild space-time

The space-time in the Solar System can be described by a static solution with spherical symmetry of the form

$$ds^2 = -e^{\mu(r)}dt^2 + e^{\nu(r)}dr^2 + r^2d\Omega^2, \quad (1.138)$$

with $d\Omega^2 = d\theta^2 + \sin^2\theta d\varphi^2$. The solution of the Einstein equations with a mass source M localized at the origin is the Schwarzschild solution whose explicit form is

$$ds^2 = -\left(1 - \frac{2G_N M}{c^2 r}\right)dt^2 + \frac{dr^2}{\left(1 - \frac{2G_N M}{c^2 r}\right)} + r^2d\Omega^2. \quad (1.139)$$

Most of the tests in the Solar System rely on the study of geodesics of this metric.

1.6.2.2 Extension

To be more general and to be able to test both the geometrical structure of space and the field equations, we will work in the weak-field approximation ($G_N M/c^2 r \ll 1$, which is justified since $G_N M_{\odot}/c^2 = 1.4 \times 10^3$ m for the Sun). We assume that the metric is of the form

$$ds^2 = -\left[1 - 2\frac{G_N M}{c^2 r} + 2(\beta^{PPN} - \gamma^{PPN})\frac{G_N^2 M^2}{c^4 r^2}\right]dt^2 + \left(1 + 2\gamma^{PPN}\frac{G_N M}{c^2 r}\right)dr^2 + r^2d\Omega^2. \quad (1.140)$$

From the field equations of motion of general relativity and using (1.139), we get that $\gamma^{PPN} = \beta^{PPN} = 1$, but there is a priori no reason for this to be the case within another metric theory of gravity.

1.6.2.3 Light bending

The study of null geodesics can be performed directly in the general form (1.138). By symmetry, we can choose to concentrate on the solutions that lie in the plane $\theta = \pi/2$. The trajectory then reduces to $\{t(\lambda), r(\lambda), \varphi(\lambda)\}$. Since space-time is static and has a spherical symmetry, there must be two constants of motion. The one associated with time-translation invariance is identified with the energy $E = u_t = g_{tt}dt/d\lambda$,

while the second one, associated with the angular momentum rotation invariance, is $L = u_\phi = g_{\varphi\varphi}d\varphi/d\lambda$. We therefore have

$$\frac{dt}{d\lambda} = E e^{-\mu(r)}, \quad r^2 \frac{d\varphi}{d\lambda} = L. \quad (1.141)$$

For a null geodesic, we have also the extra condition $k^\mu k_\mu = 0$, which we can write as

$$e^{-\mu(r)} \left(\frac{dt}{d\lambda} \right)^2 = e^{\nu(r)} \left(\frac{dr}{d\lambda} \right)^2 + r^2 \left(\frac{d\varphi}{d\lambda} \right)^2.$$

Using the constants (1.141), we can deduce that the equation of the trajectory is

$$\left(\frac{dr}{r^2 d\varphi} \right)^2 = \left(\frac{E}{L} \right)^2 e^{-\nu(r)-\mu(r)} - \frac{e^{-\nu(r)}}{r^2}. \quad (1.142)$$

The functions $\nu(r)$ and $\mu(r)$ can be identified with the corresponding terms in (1.140). After introducing $y = G_N M/c^2 r$ and $b = L/E$, we therefore have

$$\frac{d^2 y}{d\varphi^2} + y = 3\gamma^{PPN} y^2 + (1 - \gamma^{PPN}) \frac{G_N^2 M^2}{b^2 c^4}.$$

To second order in M/b , the solution of this equation is given by the relation $y = (G_N M/bc) \sin \varphi + (G_N M/bc) (1 + \gamma^{PPN} \cos^2 \varphi)$ so that the closest point from the central body is $r_0 \simeq b[1 - (G_N M/bc)]$ (see Fig. 1.12). In the limit $r \rightarrow \infty$, we get $\varphi = \varphi_+ = \pi + (1 + \gamma^{PPN})(G_N M/bc)$ or $\varphi = \varphi_- = -(1 + \gamma^{PPN})(G_N M/bc)$. The bending angle, $\Delta\varphi = \varphi_+ - \varphi_-$, is then

$$\Delta\varphi = 2(1 + \gamma^{PPN}) \frac{G_N M_\odot}{r_0 c^2}, \quad (1.143)$$

which reduces to $\Delta\varphi = 1.75''(1 + \gamma^{PPN})/2$ when $r_0 = R_\odot$.

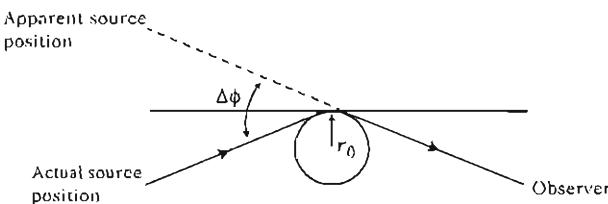


Fig. 1.12 Geometry of the trajectory of a photon deflected by a massive body.

1.6.2.4 Shapiro Effect

The Shapiro effect, discovered in 1964, is an abnormal time delay when electromagnetic waves are emitted from Earth and reflected back by a satellite or a planet. Using the null geodesic equations derived in the previous section, we obtain

$$\left(\frac{dr}{dt} \right)^2 = e^{\mu(r)-\nu(r)} \left[1 - e^{\mu(r)} \frac{b^2}{r^2} \right],$$

the coordinate time a photon takes to go from r to r_0 is therefore

$$t(r, r_0) = \int_r^{r_0} \sqrt{\frac{e^{\nu(r)} - e^{\mu(r)}}{1 - e^{\mu(r)} b^2/r^2}} dr.$$

Integrating this expression from r_{Earth} to r_{planet} , back and forth, we get that the rate of change of the travel proper time

$$\frac{d\Delta T}{d\tau} \simeq -(1 + \gamma^{\text{PPN}}) \mu\text{s} \cdot \text{h}^{-1} \quad (1.144)$$

assuming $r_0 \simeq 7 \times 10^8$ m.

1.6.2.5 Precession of the perihelion of Mercury

The planet trajectories are obtained using the time-like geodesics, such that

$$e^{-\mu(r)} E^2 = 1 + e^{\nu(r)} \left(\frac{L}{r^2} \right)^2 \left(\frac{dr}{d\varphi} \right)^2 + \left(\frac{L}{r^2} \right).$$

Introducing $y = 1/r$ and neglecting terms of order $(G_N M / rc^2)^2$, we can integrate this equation to obtain the approximate solution

$$y(\varphi) = \frac{G_N M}{L} \left(1 + e \cos \left\{ \left[1 - (2 + 2\gamma^{\text{PPN}} - \beta^{\text{PPN}}) \frac{G_N^2 M^2}{L^2} \right] \varphi \right\} \right),$$

where e is the eccentricity. At each orbit, the perihelion advances by

$$\Delta\varphi = 2\pi(2 + 2\gamma^{\text{PPN}} - \beta^{\text{PPN}}) \frac{G_N^2 M^2}{L^2} = \frac{2\pi G_N M}{a(1 - e^2)} (2 + 2\gamma^{\text{PPN}} - \beta^{\text{PPN}}). \quad (1.145)$$

For Mercury, we get $42.98''$ per century if $\gamma^{\text{PPN}} = \beta^{\text{PPN}} = 1$, which is consistent with observations.

1.6.2.6 Summary

The tests performed in the laboratory and in the Solar System give a bound to the violation of the equivalence principle and to the deviations from the theory of general relativity. We have only presented here some of these parameters, but the *parameterized post-Newtonian* (PPN) formalism brings into play other parameters that go beyond the purely Newtonian theory. This formalism has been developed in order to take into account all kinds of possible deviations. The following table and Fig. 1.13 give a summary of the main constraints obtained in the Solar System (see Ref. [12] for more details).

1.6.3 Gravitational radiation

Just as any charged object emits electromagnetic waves, a massive body can radiate gravitational waves. In what follows, we review the properties of this radiation, without giving any proof. The existence and the properties of gravitational waves can be used to test the theory of general relativity.

Table 1.1 Solar-system constraints on the post-Newtonian parameters.

| Parameters | Constraints | Details | Reference |
|---------------------------------|----------------------------------|-----------------------------------|---------------|
| η (weak) | 10^{-3} | Pendulum | Newton (1686) |
| | 5×10^{-9} | Torsion balance | [14] |
| | $(-1.9 \pm 2.5) \times 10^{-12}$ | Torsion balance (Be-Cu) | [19] |
| | 5.5×10^{-13} | Torsion balance (Earth-Moon) | [20] |
| η (strong) | $(1 \pm 15) \times 10^{-3}$ | LLR | [21] |
| α | 10^{-2} | Photon emitted (Mossbauer effect) | [22] |
| | 2×10^{-4} | H-Maser in orbit | [23] |
| δm_1^{ij} δ^2 | $1.7 \times 10^{-16} \text{ eV}$ | Lithium | [15] |
| | 10^{-23} | Strong interaction | |
| | 10^{-22} | Electromagnetic interaction | |
| | 5×10^{-18} | Weak interaction | |

1.6.3.1 Radiation of gravitational waves

The gravitational radiation emitted by an object is given by

$$h_{ij}^{(TT)}(x, t) = \frac{2G_N}{c^4 r} P_{ijkl} \frac{d^2 Q_{kl}}{dt^2} \left(t - \frac{r}{c} \right),$$

where $P_{ijkl}(n) = (\delta_{ik} - n_i n_k)(\delta_{jl} - n_j n_l) - \frac{1}{2}(\delta_{ij} - n_i n_j)(\delta_{kl} - n_k n_l)$ and where Q_{kl} is the quadrupole of the mass distribution

$$Q_{kl}(t) = \int \rho(x, t) \left(x_k x_l - \frac{1}{3} x^2 \delta_{kl} \right) d^3 x.$$

The radiation is therefore sourced by the quadrupole of the matter distribution, and there is no monopole contribution (due to the mass conservation) or dipole (from the equivalence principle). The total power of the radiation is given by the Einstein quadrupole formula

$$\frac{dE}{dt} = \frac{G_N}{5c^5} \frac{d^3 Q_{kl}}{dt^3} \frac{d^3 Q_{kl}}{dt^3}.$$

1.6.3.2 Binary pulsar PSR 1913+16

Discovered in 1974 by Taylor and Hulse, this pulsar made it possible to test the formula for the gravitational radiation and prove the existence of gravitational radiation. Using Kepler's law, it can be shown that if the energy of the binary system decreases, its period must then increase as

$$\left\langle \frac{dP}{dt} \right\rangle = -\frac{192\pi}{5c^2} \left(\frac{2\pi G_N}{P} \right)^{5/3} \frac{m_1 m_2}{m_1 + m_2} \frac{1 + 73e^2/24 + 37e^4/96}{(1 - e^2)^{7/2}},$$

where m_1 and m_2 are the two masses and e is the eccentricity. For the pulsar PSR 1913+16, $m_1 = 1.44 M_\odot$ and $m_2 = 1.38 M_\odot$, $e = 0.617$ and $P = 7 \text{ h } 40 \text{ min}$ so that

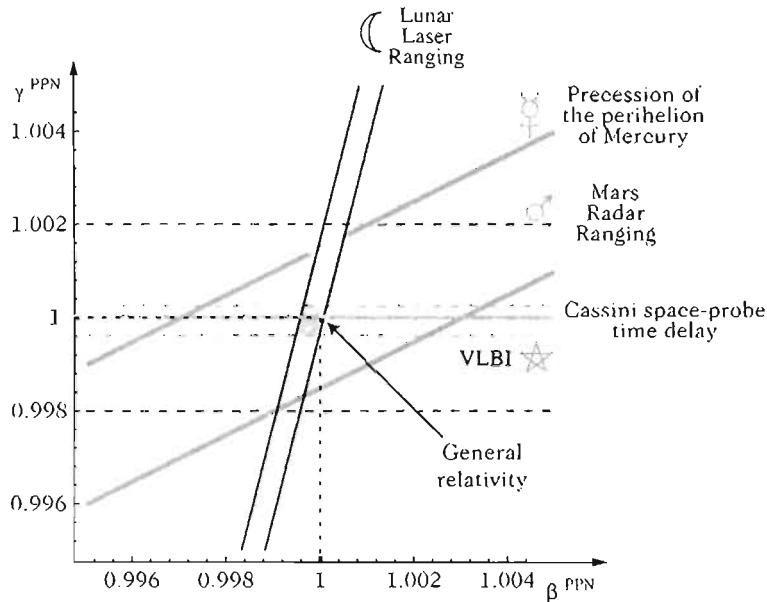


Fig. 1.13 Summary of the constraints on the post-Newtonian parameters in the Solar System. The current constraints are $|2\gamma^{\text{PPN}} - \beta^{\text{PPN}} - 1| < 3 \times 10^{-3}$ from the precession of the perihelion of Mercury, $4\beta^{\text{PPN}} - \gamma^{\text{PPN}} - 3 = (-0.7 \pm 1) \times 10^{-3}$ from the Lunar laser ranging (LLR) experiment, $|\gamma^{\text{PPN}} - 1| < 4 \times 10^{-4}$ from the light bending (VLBI: Very Long Baseline Interferometer) and $|\gamma^{\text{PPN}} - 1| = (2.1 \pm 2.3) \times 10^{-5}$ from the measure of the time delay of the Cassini space probe signal.

$\langle dP/dt \rangle = -2.4 \times 10^{-12}$. We may note that there is, however, a real discrepancy of 14σ between the observed decrease and the theoretical prediction. This is because the Doppler effect coming from the acceleration of the pulsar towards the centre of the Galaxy should be taken into account. The theoretical works in Ref. [24] have made it possible to compute the equations of motion up to order v^5/c^5 and have led to ideas on how to test general relativity in the strong-field limit [25] (see Ref. [26] for a review). They also prove that gravity propagates at the speed of light.

1.6.3.3 Direct detection

Many experiments are now in progress to detect gravitational waves, either from interferometry (LIGO, VIRGO, LISA, ...) or using Weber bars (see Ref. [27] for a review).

In order to understand how difficult it is to detect them, let us give an estimate of the amplitude of the gravitational waves. Consider for that a bar of length L and mass M rotating with angular velocity ω . The characteristic amplitude of the quadrupole is $Q = ML^2$, so that $d^n Q/dt^n \sim ML^2\omega^n$. We therefore have a typical amplitude

$$h \sim \frac{G_N M L^2 \omega^2}{c^4 r}, \quad \frac{dE}{dt} \sim \frac{G_N M^2 L^4 \omega^6}{c^5}.$$

For a 500-ton mass and 20-m long bar rotating at 5 rad/s, we get $h \sim 10^{-38}$ at 50 m, and $dE/dt \sim 10^{-32}$ W! The radiation of ordinary objects is therefore undetectable. For a binary system of typical mass $1M_\odot$, with an orbit of 10⁶ km and an orbital period of order 10 h, we get $dE/dt \sim 10^{22}$ W. This power can even be increased and reach a typical value of $dE/dt \sim c^5/G_N \sim 3.6 \times 10^{52}$ W for an ultrarelativistic source.

1.6.4 Need for tests at astrophysical scales

Most of the tests presented here have a characteristic scale of the order of the size of the Solar System or at most, of the size of our Galaxy for binary pulsars. In cosmology, we will extrapolate the law of gravitation to the Universe itself. This extrapolation requires some checks that we will present in detail in the rest of this book. Let us, nevertheless, remind ourselves of some facts worth keeping in mind (see Ref. [28] for more details):

1. At galaxy scales, the rotation curves indicate that the gravitational potential goes as $\phi \propto \ln r$ on large scales. This behaviour is usually explained by the presence of dark matter, which has not been detected yet. Beyond a given scale, everything goes effectively as if gravity had a bidimensional behaviour. The study of the rotation curves, however, shows that if such a transition existed, this scale should depend on the galaxy mass (see Chapter 7 for more details on these relations).
2. For galaxy clusters of around 2 Mpc, the effects of gravitational lensing, which measures the gravitational potential from light bending, and the X-ray emission, which traces out the gas temperature, can be compared. This comparison has shown that the Poisson equation is valid up to a factor 2. Hence, it provides a test of general relativity (testing both the light bending and Einstein's equations in the weak-field limit).
3. At astrophysical and cosmological scales, there are no direct tests of the law of gravity. The growth of the large-scale structure of the Universe provides a test, but it mixes the properties of gravity with those of matter. For the theory to be consistent with the observations, dark matter should be present.
4. The observed recent accelerated expansion of the Universe (see Chapter 4) cannot be explained within the framework of general relativity with ordinary matter (among other conditions, matter with positive pressure; this follows from the Raychaudhuri equation). This observation is explained either through exotic matter (some candidates such as the cosmological constant, the quintessence,... will be described in Chapter 12) or by modifying the laws of gravity at large scales.
5. Two tests are currently proposed. On the one hand, the Einstein equivalence principle can be tested by checking how constant the 'constants of Nature' are on large scales. On the other hand, the Poisson law can be tested up to scales of a hundred megaparsecs, by comparing the matter distribution in the galaxy catalogues with the weak lensing by large-scale structures (see Ref. [29]).

This shows the importance of testing gravity either in order to validate the standard approach that adds new exotic matter sectors (dark matter, dark energy), or in order to give new explanations for the Universe. We will come back to all these issues in Chapter 12.

References

- [1] R. HAKIM, *An introduction to relativistic gravitation*, Cambridge University Press, 1999.
- [2] H. STEPHANI, *General relativity*, Cambridge University Press, 1990.
- [3] S. WEINBERG, *Gravitation and cosmology*, John Wiley and Sons, 1972.
- [4] R. WALD, *General relativity*, Chicago University Press, 1984.
- [5] L. LANDAU and E. LIFSHITZ, *Course of theoretical physics*, Volume 2, Pergamon Press, 1976.
- [6] J.D. JACKSON, *Classical electrodynamics*, Wiley, 1998.
- [7] J. EISENSTAEDT, *Einstein et la relativité générale*, CNRS Éditions, 2003.
- [8] C.M. WILL, *Theory and experiments in gravitational physics*, Cambridge University Press, 1984.
- [9] G.F.R. ELLIS, ‘Relativistic cosmology’, in *Proc. Intl. School of Physics ‘Enrico Fermi’*, Corso XLVII, R.K. Sachs (ed.), pp. 104–179, Academic Press, 1971.
- [10] S.W. HAWKING and G.F.R. ELLIS, *The large scale structure of space-time*, Cambridge University Press, 1973.
- [11] J. BARDEEN, ‘Gauge-invariant cosmological perturbations’, *Phys. Rev. D* **22**, 1882, 1981.
- [12] C.M. WILL, ‘The confrontation between general relativity and experiment’, *Living Rev. Relativity* **4**, 4, 2001.
- [13] I. CIUFOLINI and J.A. WHEELER, *Gravitation and inertia*, Princeton University Press, 1995.
- [14] R.V. EÖTVÖS, V. PEKÁR and E. FEKETE, ‘Beitrage zum Gesetze der Proportionalität von Trägheit und Gravität’, *Ann. Physik* **68**, 11, 1922.
- [15] V.W. HUGHES *et al.*, ‘Upper limit for the anisotropy of inertial mass from nuclear resonance experiments’, *Phys. Rev. Lett.* **4**, 342, 1960; R.W. DREVER, ‘A search for the anisotropy of inertial mass using a free precession technique’, *Phil. Mag.* **6**, 683, 1961.
- [16] J.P. UZAN, ‘The fundamental constants and their variation: observational and theoretical status’, *Rev. Mod. Phys.* **75**, 403, 2003.
- [17] E. FISCHBACH and C. TALMADGE, ‘Ten years of the fifth force’, in *Dark matter in cosmology, quantum measurements, experimental gravitation*, Moriond proceedings, 443 (1996), arXiv:hep-ph/9606249; E. ADELBERGER, B. HECKEL, and A. NELSON, ‘Tests of the gravitational inverse square law’, *Ann. Rev. Nuc. Phys.* **53**, 77, 2003.
- [18] C.D. HOYLE *et al.*, ‘Sub-millimeter tests of the gravitational inverse square law’, *Phys. Rev. D* **70**, 042004, 2004.
- [19] Y. SU *et al.*, ‘New tests of the universality of free fall’, *Phys. Rev. D* **50**, 3614, 1994.

- [20] S. BAESSLER *et al.*, ‘Improved test of the equivalence principle for gravitational self-energy’, *Phys. Rev. Lett.* **83**, 3585, 1999.
- [21] J.O. DICKEY *et al.*, ‘Lunar laser ranging: a continuous legacy of the Apollo program’, *Science* **265**, 482, 1994.
- [22] R.V. POUND and G.A. REBKA, ‘Apparent weight of photons’, *Phys. Rev. Lett.* **4**, 337, 1960.
- [23] R. VESSOT and M. LEVINE, ‘A test of the equivalence principle using a space-borne clock’, *J. Gen. Rel. Grav.* **10**, 181, 1979.
- [24] T. DAMOUR and N. DERUELLE, ‘General relativistic celestial mechanics of binary systems’, *Ann. Inst. H. Poincaré* **43**, 107, 1985; **44**, 263, 1986.
- [25] T. DAMOUR and J.H. TAYLOR, ‘Strong field tests of relativistic gravity and binary pulsars’, *Phys. Rev. D* **45**, 1840, 1992.
- [26] I.H. STAIRS, ‘Testing relativity with pulsar timing’, *Living Rev. Relativity* **5**, 1, 2003.
- [27] I. BLAIR, ‘Detection of gravitational waves’, *Rep. Prog. Phys.* **63**, 1317, 2000.
- [28] J.-P. UZAN, ‘Tests of gravity on large scales and variation of the constants’, *Int. J. Theor. Phys.* **42**, 1163, 2003.
- [29] J.-P. UZAN and F. BERNARDEAU, ‘Lensing at cosmological scales: a test of higher dimensional gravity’, *Phys. Rev. D* **64**, 083004, 2000.

2

Overview of particle physics and the Standard Model

Although gravity structures space-time on large scales, numerous physical processes rely on the description of the three other interactions, namely electromagnetism, the weak and strong nuclear forces. This chapter is dedicated to the description of these interactions and of the particles on which they act. This *standard model of particle physics* describes ordinary matter, as observed in the laboratory. We shall see later that the cosmological observations concerning the existence of dark matter call for an extension of this framework. It is hence necessary to present it in detail. For this, we will work in this chapter in a flat space-time.

Section 2.1 recalls the basis of analytical mechanics and the steps from classical mechanics to quantum mechanics. Section 2.2 then focuses on the special case of a scalar field. Being the simplest prototypal model, the free or self-interacting scalar field illustrates in a complete way what can be learnt from quantum field theory.

The following sections sketch a general picture of the standard model of particle physics. Section 2.3 presents the spectrum of particles detected in accelerators, and Section 2.4 describes how, relying on the notion of symmetry, a classification can be sketched out. The Higgs mechanism will then be described in Section 2.5. Applying this mechanism to electroweak interactions will allow us to obtain the standard model of particle physics, presented in Section 2.6. Finally, Section 2.7 presents a study of discrete symmetries, followed by an overview on the current state of this standard model.

The ideas introduced in this chapter will be used repeatedly in the rest of the book. For instance, the aspects of quantization and in particular those of scalar fields will be central to our discussion of inflation (Chapter 8), and also when we touch on the cosmological constant problem (Chapter 12); the notion of symmetry breaking plays a crucial role in the formation of topological defects (Chapter 11) and in the grand unified theories. This chapter therefore lays the necessary basis for the study of cosmologically relevant topics such as inflation of course, but also of topological defects, the origin of dark matter, baryogenesis, the cosmological constant problem, etc.

2.1 From classical to quantum

It is now a little more than a century since the principles of quantum mechanics started being used to describe the laws of Nature. This section summarizes the steps from analytical classical mechanics to quantum mechanics.

2.1.1 Analytical classical mechanics

Formally, quantum mechanics can be seen as an extension of classical mechanics as formalized finally, e.g., by Lagrange in the second half of the eighteenth century and Hamilton at the beginning of the nineteenth century. The classical laws can be transformed into their quantum counterparts by applying the *correspondence principle*.

2.1.1.1 Analytical classical mechanics

The essence of classical mechanics (all the details of which can be found in numerous works, in Refs. [1,2] among others) can be summarized by the existence of a Lagrangian function L from which all the dynamics of a system can be derived. For a set of N interacting particles, this Lagrangian depends on the coordinates x_i , ($i = 1, \dots, N$) of these particles, and on their velocities, \dot{x}_i . We merge together the set of all these coordinates, which are not necessarily Cartesian, under the designation q_i (*generalized coordinates*), with the index i now varying from 1 to $3N$, and the generalized velocities \dot{q}_i . The dynamics of this system can be obtained by minimizing the action

$$S = \int_{t_{\text{initial}}}^{t_{\text{final}}} L[q_i(t), \dot{q}_i(t), t] dt, \quad t_{\text{initial}}, t_{\text{final}} = \text{const.}, \quad (2.1)$$

assuming that the value of the coordinates is fixed at the boundary of the integration domain [$\delta S = 0$ with $\delta q_i(t_{\text{initial}}) = \delta q_i(t_{\text{final}}) = 0$]. The Euler–Lagrange equations are then obtained as

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \right) = \frac{\partial L}{\partial q_i}. \quad (2.2)$$

The first term of this equation brings into play the quantities $p^i \equiv \partial L / \partial \dot{q}_i$, which are the *generalized momenta*.

It is possible to invert these relations to determine the generalized velocities in terms of the coordinates and momenta. Thanks to a Legendre transformation, all the dynamics can then be described by changing variables and using the generalized momenta instead of the velocities. For that, let us define the Hamilton function, or Hamiltonian H , by

$$H[q_i(t), p^i(t), t] \equiv \sum_{i=1}^{3N} p^i \dot{q}_i - L\{q_i(t), \dot{q}_i[p^j(t)], t\}, \quad (2.3)$$

which allows us to write the second-order differential equations (2.2) as a first-order system,

$$\dot{p}^i = -\frac{\partial H}{\partial q_i}, \quad \dot{q}_i = \frac{\partial H}{\partial p^i}. \quad (2.4)$$

This system allows us to easily obtain the time derivative of any physical quantity by using the *Poisson brackets* defined by

$$\{A, B\} \equiv \sum_i \left(\frac{\partial A}{\partial q_i} \frac{\partial B}{\partial p^i} - \frac{\partial A}{\partial p^i} \frac{\partial B}{\partial q_i} \right), \quad (2.5)$$

A and B being two quantities depending on q_i and p^i . Substituting for B the Hamiltonian H in the definition (2.5), and using the equations of motion (2.4), the total time derivative of A can easily be seen to be

$$\frac{dA}{dt} = \frac{\partial A}{\partial t} + \sum_i \left(\frac{\partial A}{\partial p^i} \dot{p}^i + \frac{\partial A}{\partial q_i} \dot{q}_i \right) = \frac{\partial A}{\partial t} + \{A, H\}. \quad (2.6)$$

Finally, the Poisson brackets of the coordinates and momenta satisfy

$\{q_i, q_j\} = 0 = \{p^i, p^j\} \quad \text{and} \quad \{q_i, p^j\} = \delta_i^j.$

(2.7)

This formalism introduces an antisymmetric operator between two quantities, i.e. such that $\{A, B\} = -\{B, A\}$. This property will allow us to make the transition formally but very easily to the quantum description.

2.1.1.2 Functional derivative

Classical field theory, which can be illustrated, for instance, by general relativity, can be written using the same formalism as the analytical theory of point particles, if we rely on field functionals rather than on functions.

A functional $Z[f]$, is defined as a map from a normalized linear space of functions,¹ $\mathcal{L} = \{f(x); x \in \mathbb{R}\}$, onto the set of complex (or real in some cases) numbers, i.e. $Z : \mathcal{L} \mapsto \mathbb{C}$ (or \mathbb{R}). The functional derivative is then defined in a similar way as the partial derivative for a function of many variables. A functional can indeed be understood as a function with an infinite number of variables. Its infinitesimal variation is

$$\delta Z[f] = \int \frac{\delta Z[f]}{\delta f(x)} \delta f(x) dx, \quad (2.8)$$

introducing the dummy variable x , this reduces to a definition similar to the one in normal analysis, that is

$$\frac{\delta Z[f(x)]}{\delta f(y)} \equiv \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \{Z[f(x) + \epsilon \delta(x - y)] - Z[f(x)]\}, \quad (2.9)$$

where δ is the Dirac distribution in one dimension. Note that for all computations, the functions are supposed to be integrable and to vanish fast enough at large x in order for the surface terms to vanish in the integration by parts.

The two definitions (2.8) and (2.9) are equivalent and can both easily be shown to satisfy the rules required for a derivative, i.e. the Leibniz rule

$$\frac{\delta}{\delta f(x)} (Z[f]W[f]) = \frac{\delta Z[f]}{\delta f(x)} W[f] + Z[f] \frac{\delta W[f]}{\delta f(x)}, \quad (2.10)$$

and the functional composition law

¹Technically, this is called a Banach space.

$$\frac{\delta Z\{W[f]\}}{\delta f(x)} = \int \frac{\delta Z[W]}{\delta W(y)} \frac{\delta W[f]}{\delta f(x)} dy. \quad (2.11)$$

From the definition (2.9), it can easily be shown that

$$\frac{\delta f(x)}{\delta f(y)} = \delta(x - y), \quad (2.12)$$

where the first function, $f(x)$, is seen here as the trivial functional that maps any function into itself. We also have the useful property

$$\frac{\delta}{\delta f(y)} \int g \left[\frac{df(x)}{dx} \right] dx = - \frac{d}{dy} g' \left[\frac{df(y)}{dy} \right],$$

where the prime indicates a simple derivative of the function with respect to its argument.

Finally, it is interesting to show that, from the functional point of view, a function and its derivative are not independent. By choosing $Z[f(x)] = f'(x)$, and using the definition (2.8), we find that

$$\delta f'(y) = \int \frac{\delta f'(y)}{\delta f(x)} \delta f(x) dx = \int \frac{d}{dy} \left[\frac{\delta f(y)}{\delta f(x)} \right] \delta f(x) dx = \int \frac{d}{dy} [\delta(y - x)] \delta f(x) dx,$$

and we can hence make the identification

$$\frac{\delta f'(y)}{\delta f(x)} = \frac{d}{dy} [\delta(y - x)].$$

All of these relations can easily be generalized to an arbitrary number of dimensions. Most importantly, they allow us to write a classical field theory in a manner that bears close resemblance to the Hamiltonian formulation of a classical particle.

2.1.1.3 Lagrangian of a classical field

The equivalent of the action (2.1) for a scalar field is obtained from a Lagrangian density $\mathcal{L}[\phi(x^\mu)]$, as

$$S_{\text{field}} = \int L(t) dt = \int dt \int d^3x \mathcal{L}[\phi(x, t), \partial_\mu \phi(x, t)]. \quad (2.13)$$

Let us first consider \mathcal{L} as a function of x and t . Using an ordinary derivative followed by an integration by parts in four dimensions gives rise to the Euler–Lagrange equations, equivalent to (2.2) for a scalar field,

$$\boxed{\frac{\partial \mathcal{L}}{\partial \phi(x^\alpha)} = \frac{\partial}{\partial x^\mu} \frac{\partial \mathcal{L}}{\partial |\partial_\mu \phi(x^\alpha)|}}. \quad (2.14)$$

This relation can also be obtained using a functional differentiation, i.e. viewing \mathcal{L} as a functional of the field and its time derivative, in the following way. The time and spatial dependencies of the Lagrangian can be separated explicitly before the variation

$$\delta L = \int d^3x \left(\left\{ \frac{\partial \mathcal{L}}{\partial \phi(x,t)} - \nabla \frac{\partial \mathcal{L}}{\partial [\nabla \phi(x,t)]} \right\} \delta \phi(x,t) + \frac{\partial \mathcal{L}}{\partial \dot{\phi}(x,t)} \delta \dot{\phi}(x,t) \right),$$

where an integration by parts was performed to get the second term. The definition (2.8) for the functional derivative then implies that

$$\delta L = \int d^3x \left[\frac{\delta \mathcal{L}}{\delta \phi(x,t)} \delta \phi(x,t) + \frac{\delta \mathcal{L}}{\delta \dot{\phi}(x,t)} \delta \dot{\phi}(x,t) \right],$$

and therefore, by identification, we obtain

$$\frac{\delta \mathcal{L}}{\delta \phi(x,t)} = \frac{\partial \mathcal{L}}{\partial \phi(x,t)} - \nabla \frac{\partial \mathcal{L}}{\partial [\nabla \phi(x,t)]}, \quad \text{and} \quad \frac{\delta \mathcal{L}}{\delta \dot{\phi}(x,t)} = \frac{\partial \mathcal{L}}{\partial \dot{\phi}(x,t)}.$$

Integrating these relations by parts, assuming that the variation at the boundary vanishes, this reduces to (2.14).

2.1.1.4 Hamiltonian and field Poisson brackets

Starting from the Lagrangian (2.13), and using an approach completely similar to that used in point-particle mechanics, we can first define the canonically conjugate field $\pi(x,t)$, which is equivalent to the generalized momentum [see (2.2)], by

$$\pi(x,t) \equiv \frac{\delta L(t)}{\delta \dot{\phi}(x,t)}, \quad (2.15)$$

so that the equation of motion is of the same form as for point particles, namely $\dot{\pi}(x,t) = \delta L / \delta \phi(x,t)$.

In order to generalize (2.3) to the scalar field case, the Hamiltonian can be written in the following form

$$H(t) = \int \pi(x,t) \dot{\phi}(x,t) d^3x - L(t) \equiv \int \mathcal{H}(x,t) d^3x, \quad (2.16)$$

which hence defines the Hamiltonian density, related to the Lagrangian density \mathcal{L} , by the relation

$$\mathcal{H}(x,t) = \pi(x,t) \dot{\phi}(x,t) - \mathcal{L}(x,t). \quad (2.17)$$

In this formalism, the Euler–Lagrange equations (2.14) get replaced by the set of Hamilton equations

$$\dot{\phi} = \frac{\delta H}{\delta \pi}, \quad \dot{\pi} = -\frac{\delta H}{\delta \phi}.$$

(2.18)

This system of first-order equations is the equivalent of (2.4) for a scalar field.

The Poisson bracket of two functionals is constructed in the same way as for two classical quantities via the straightforward generalization of (2.5)

$$\{W[\phi(x), \pi(x)], Z[\phi(x), \pi(x)]\} \equiv \int d^3x \left[\frac{\delta W}{\delta \phi(x)} \frac{\delta Z}{\delta \pi(x)} - \frac{\delta W}{\delta \pi(x)} \frac{\delta Z}{\delta \phi(x)} \right]. \quad (2.19)$$

The Hamilton equations can hence provide the time evolution of any functional W by

$$\frac{dW}{dt} = \frac{\partial W}{\partial t} + \{W, H\}.$$

The conjugate momentum π and the field ϕ obey the relations

$$\begin{aligned} \{\phi(x, t), \phi(x', t)\} &= \{\pi(x, t), \pi(x', t)\} = 0, \\ \{\phi(x, t), \pi(x', t)\} &= \delta^{(3)}(x - x'), \end{aligned} \quad (2.20)$$

which generalizes (2.7).

2.1.1.5 Noether Theorem and conservation laws

In 1918, E. Noether enunciated the theorem according to which a conserved quantity can be associated to any transformation that leaves the action of a theory invariant. More precisely, this theorem is enunciated in the following way:

If the action of a given theory is invariant under the infinitesimal transformations of a symmetry group, then there exist as many conserved quantities (constant in time) as infinitesimal transformations associated to this symmetry group.

Furthermore, these conserved quantities can be obtained from the Lagrangian density.

As an example of this theorem, as well as for its own interest, let us mention the conservation of energy that results from the time translation invariance of the laws of physics (one invariance, one conserved quantity), as well as the conservation of momentum and angular momentum, which arise, respectively, from the translation and the rotation invariances of space-time. In these two last cases, the invariance is vector-like (three coordinates for the translation vector, or three angles for the rotation), and the conserved quantity remains of the same nature. All these invariances, which are usual in classical mechanics, follow from the conservation of a unique quantity, the energy-momentum tensor. This construction can be generalized to the case of general relativity (Chapter 1, Section 1.4).

2.1.1.6 Case of classical point particles

The classical mechanics of point particles is completely described by specifying a Lagrangian $L(q_i, \dot{q}_i, t)$. If the physics that corresponds to this Lagrangian is time-translation invariant, i.e. such that the result of any experiment involving this theory does not depend on the time at which it is made, it is then expected that

$L(q_i, \dot{q}_i, t - t_0) = L(q_i, \dot{q}_i, t)$, for any arbitrary initial time t_0 . This implies that the Lagrangian cannot explicitly depend on time, i.e. $L(q_i, \dot{q}_i, t) = L(q_i, \dot{q}_i)$; it depends on time only through the variables q_i . It follows that $\partial L / \partial t = 0$, so that its total derivative with respect to time is

$$\frac{dL}{dt} = \sum_i \left(\frac{\partial L}{\partial q_i} \frac{dq_i}{dt} + \frac{\partial L}{\partial \dot{q}_i} \frac{d\dot{q}_i}{dt} \right). \quad (2.21)$$

The equations of motion (2.2) then imply that the function E , defined as

$$E \equiv \sum_i \dot{q}_i \frac{\partial L}{\partial \dot{q}_i} - L, \quad (2.22)$$

is conserved in time, i.e. $dE/dt = 0$. One can easily be convinced that E is indeed an energy if we take the example of a set of N point particles of mass m in a potential V for which

$$L_{\text{ex}} = \sum_{i=1}^{3N} \frac{1}{2} m \dot{q}_i^2 - V(q_j). \quad (2.23)$$

The relation (2.22) then implies that

$$E_{\text{ex}} = \sum_i \frac{1}{2} m \dot{q}_i^2 + V(q_j),$$

which corresponds to the classical expression for the total energy (kinetic plus potential) of a system of point particles. Note that (2.22) is simply the Hamiltonian (2.3).

The invariance under infinitesimal coordinates transformations and generalized velocities, is obtained using the transformation $q_i \mapsto Q_i = q_i + \epsilon f_i$, and $\dot{q}_i \mapsto \dot{Q}_i = \dot{q}_i + \epsilon \dot{f}_i$, where ϵ is a small parameter ($\epsilon \ll 1$) and f_i is a vector that will be specified later. The invariance under this transformation can be written as

$$L(Q_i, \dot{Q}_j) = L(q_i, \dot{q}_j).$$

Expanding the left-hand side of this relation to leading order in ϵ , we get

$$\sum_i \left(f_i \frac{\partial L}{\partial q_i} + \dot{f}_i \frac{\partial L}{\partial \dot{q}_i} \right) = 0.$$

Once again, the equations of motion (2.2) imply that

$$\frac{d}{dt} \left(\sum_i f_i \frac{\partial L}{\partial \dot{q}_i} \right) = 0, \quad (2.24)$$

and hence to the existence of a new conserved quantity.

Equation (2.24) can be illustrated on the example (2.23) if one chooses translation in space as the coordinate transformation (i.e. $f_i = \ell_i$ constants in Cartesian coordinates). For each of the three possible independent quantities ℓ_i (i.e. along the three

possible axes), the quantity $p = m\dot{q}$ is conserved: this is the momentum. One can clearly see here that for a vector-like invariance (three independent quantities form a vector) corresponds to a conservation law that is itself vector-like. This is the meaning of Noether's theorem.

2.1.1.7 Field theory case

Let us now come back to the case of field theory, an example of which is the action (2.13), and let us consider a general coordinate transformation

$$x^\mu \mapsto x'^\mu = x^\mu + \epsilon f^\mu. \quad (2.25)$$

Then the scalar field ϕ transforms as

$$\phi'(x'^\alpha) = \phi(x^\alpha) + \epsilon \delta\phi(x^\alpha),$$

which is a local variation. One can also define a global variation \bar{d} through

$$\epsilon \bar{d}\phi(x^\alpha) \equiv \phi'(x^\alpha) - \phi(x^\alpha),$$

which only takes into consideration the modification of the field itself at the point x^α , thus defining a total variation after the local variation is brought back to its original point. To first order in ϵ , these two variations are related by

$$\bar{d}\phi(x^\alpha) = \delta\phi(x^\alpha) - f^\mu \partial_\mu \phi(x^\alpha) + \mathcal{O}(\epsilon^2). \quad (2.26)$$

Note that the variation \bar{d} commutes with the derivative, unlike the variation δ . Indeed,

$$\epsilon \partial_\alpha \bar{d}\phi(x^\beta) = \partial_\alpha [\phi'(x^\beta) - \phi(x^\beta)] = \partial_\alpha \phi'(x^\beta) - \partial_\alpha \phi(x^\beta) = \epsilon \bar{d} \partial_\alpha \phi(x^\beta),$$

while

$$\partial_\alpha [\epsilon \delta\phi(x^\beta)] = \partial_\alpha [\phi'(x^\beta) + \epsilon f^\mu \partial_\mu \phi'(x^\beta) - \phi(x^\beta)] = \epsilon [\delta \partial_\alpha \phi(x^\beta) + f^\mu \partial_\alpha \partial_\mu \phi(x^\beta)],$$

still to first order in ϵ .

The theory will be said to be invariant under the transformation $x \mapsto x'$ (to simplify, we write $x \equiv x^\alpha$ when there is no ambiguity) if the action, i.e. the integration of the Lagrangian density over a space-time volume V_4 , remains unchanged during the operation. In other words,

$$S = \int_{V_4} \mathcal{L} \left[\phi(x), \frac{\partial \phi(x)}{\partial x^\mu} \right] d^4x = \int_{V'_4} \mathcal{L}' \left[\phi(x'), \frac{\partial \phi(x')}{\partial x'^\mu} \right] d^4x', \quad (2.27)$$

where the integration is performed over the same volume of space-time, expressed in terms of the new coordinates (hence the notation V'_4 in the second equality). The volume element is modified by the determinant of the Jacobian matrix

$$d^4x' = \left| \frac{\partial(x'^\mu)}{\partial(x^\nu)} \right| d^4x \simeq (1 + \epsilon \partial_\mu f^\mu) d^4x,$$

where the second equality comes from the expansion to first order in ϵ . Equation (2.27) then becomes

$$\int_{V_4} (\delta \mathcal{L} + \mathcal{L} \partial_\mu f^\mu) d^4x = \mathcal{O}(\epsilon),$$

using the relation (2.26) between the operators δ and \bar{d} , this leads to

$$\int_{V_4} [\bar{d}\mathcal{L} + \partial_\mu (f^\mu \mathcal{L})] d^4x = 0. \quad (2.28)$$

Furthermore, we can expand $\bar{d}\mathcal{L}$ in this expression as

$$\bar{d}\mathcal{L} = \frac{\partial \mathcal{L}}{\partial \phi} \bar{d}\phi + \frac{\partial \mathcal{L}}{\partial \partial_\mu \phi} \bar{d}\partial_\mu \phi,$$

which is possible since \bar{d} is an exact differential (\bar{d} and ∂_α commute). Integrating this expression and using the field equations (2.14), we get

$$\int_{V_4} d^4x \bar{d}\mathcal{L} = \int_{V_4} d^4x \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial \partial_\mu \phi} \bar{d}\phi \right),$$

such that (2.28) becomes

$$\int_{V_4} \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial \partial_\mu \phi} \bar{d}\phi + f^\mu \mathcal{L} \right) d^4x = 0.$$

Since V_4 is arbitrary, this implies that the integrand vanishes, or in other words, that the term in parenthesis is a conserved current \mathcal{J}^μ , i.e. that satisfies $\partial_\mu \mathcal{J}^\mu = 0$. This current is explicitly given, using (2.26), by

$$\mathcal{J}^\mu = \frac{\partial \mathcal{L}}{\partial \partial_\mu \phi} \delta\phi - \left(\frac{\partial \mathcal{L}}{\partial \partial_\mu \phi} \partial^\alpha \phi - \eta^{\mu\alpha} \mathcal{L} \right) f_\alpha, \quad (2.29)$$

for a transformation of the form (2.25). In this expression we have reintroduced the local variation $\delta\phi$, which is often better controlled than the total variation $\bar{d}\phi$.

The previous discussion does not depend on the fact that ϕ is a scalar, i.e. on the way it transforms under a change of coordinates, nor does it depend on the fact that we have considered only one field. More generically, for a theory with N fields ϕ_a , where $a \in [1, N]$, the relation (2.29) can be generalized to

$$\mathcal{J}^\mu = \sum_{a=1}^N \frac{\partial \mathcal{L}}{\partial \partial_\mu \phi_a} (\delta\phi_a - f_a \partial^\alpha \phi_a) + \eta^{\mu\alpha} f_\alpha \mathcal{L}. \quad (2.30)$$

This relation is still valid if the indices of ϕ_a form, for instance, a vector ($\phi_a \rightarrow A_\mu$) or a tensor of higher rank.

The conservation of the current \mathcal{J} implies the existence of a quantity constant in time. To see this, it is sufficient to evaluate the integral of $\partial_\mu \mathcal{J}^\mu$ on an arbitrary space volume V_3 , namely

$$0 = \int_{V_3} \partial_\mu \mathcal{J}^\mu d^3x = \int_{V_3} (\partial_0 \mathcal{J}^0 + \nabla \cdot \mathcal{J}) d^3x,$$

before using the theorem of Stokes–Ostrogradski to eliminate the second term

$$\int_{V_3} d^3x \nabla \cdot \mathcal{J} = \oint_{\partial V_3} dS \cdot \mathcal{J} \xrightarrow[V_3 \rightarrow \infty]{} 0.$$

It thus follows that

$$\int_{V_3} \partial_0 \mathcal{J}^0 d^3x = \frac{d}{dt} \int_{V_3} \mathcal{J}^0 d^3x = 0,$$

and the integral of the time component of the current is indeed conserved (see the generalization of Section 1.4 from Chapter 1).

2.1.1.8 Field energy and momentum

Just as time and space translation invariance imply the conservation of energy and momentum in point-particle mechanics, the same invariance, expressed in terms of the current (2.30), generates the conservation of the energy-momentum tensor. Under an infinitesimal transformation $x'^\mu = x^\mu + \epsilon \ell^\mu$ in Minkowskian coordinates, where the ℓ^μ are four constant quantities, we can see that the field remains unchanged, i.e. $\delta\phi = 0$. Note that this remains true for any vector-like or tensor quantity of arbitrary rank, since $\partial x'^\mu / \partial x^\nu = \delta_\nu^\mu$ for this specific case.

Since the quantities ℓ_α are supposed to be constant, it is possible to read them off from the definition (2.30). The conserved quantity is then a tensor of rank 2, since

$$\partial_\mu \left(-\frac{\partial \mathcal{L}}{\partial \partial_\mu \phi} \partial^\alpha \phi + \eta^{\mu\alpha} \mathcal{L} \right) \epsilon \ell_\alpha = 0,$$

which allows us to define a canonical energy-momentum tensor $\Theta^{\mu\alpha}$

$$\Theta^{\mu\alpha} = \eta^{\mu\alpha} \mathcal{L} - \frac{\partial \mathcal{L}}{\partial \partial_\mu \phi} \partial^\alpha \phi.$$

(2.31)

This tensor is conserved by construction, i.e.

$$\partial_\mu \Theta^{\mu\alpha} = 0. \quad (2.32)$$

For each value of $\alpha = 0, \dots, 3$, corresponding to independent variations of ℓ_α , is associated a conserved quantity: they are the components of the energy-momentum vector $P^\alpha = (E/c, \mathbf{P})$, given by

$$P^\alpha = \frac{1}{c} \int_{V_3} d^3x \Theta^{0\alpha}(x^\beta), \quad (2.33)$$

which is hence conserved in time.

2.1.1.9 Symmetrization of the energy-momentum tensor

The energy-momentum tensor (2.31) is not necessarily symmetric and thus does not necessarily correspond to the tensor (1.83) that appears in the Einstein equations (1.85). However, it is always possible to make it symmetric in an appropriate way, such that it remains conserved and such that the physical quantities we can derive from it are not affected. This allows us to use its symmetric version as the source for the Einstein equations.

Let us define

$$T^{\mu\alpha} \equiv \Theta^{\mu\alpha} + \partial_\rho G^{\rho\mu\alpha}.$$

This tensor is conserved by construction if $G^{\rho\mu\alpha}$ is antisymmetric in its first two indices, i.e. $G^{\mu\rho\alpha} = -G^{\rho\mu\alpha}$. Furthermore, $T^{\mu\alpha}$ leads to the same conserved quantity as long as the fields decrease fast enough at infinity. We indeed have

$$\begin{aligned} P^\alpha &= \frac{1}{c} \int_{V_3} d^3x (T^{0\alpha} - \partial_\rho G^{00\alpha}) \\ &= \frac{1}{c} \int_{V_3} d^3x T^{0\alpha} - \frac{1}{c} \frac{d}{dt} \int d^3x G^{00\alpha} - \frac{1}{c} \int d^3x \partial_\rho G^{00\alpha}. \end{aligned} \quad (2.34)$$

The second term of the last equality vanishes identically since $G^{\rho\mu\alpha}$ is antisymmetric. The last term can be expressed as an integral over the surface that limits V_3 , i.e. ∂V_3 , which we send to infinity, where the fields are supposed to vanish. It follows that

$$P^\alpha = \frac{1}{c} \int_{V_3} d^3x T^{0\alpha}, \quad (2.35)$$

which we can identify with the momentum vector.

This can be illustrated by the example of free electromagnetism. In that case, the field ϕ gets replaced by the vector potential A_μ . Its dynamics is determined by the Lagrangian density [cf. (1.94)]

$$\mathcal{L}_{e.m.} = -\frac{1}{4} F_{\alpha\beta} F^{\alpha\beta}, \quad (2.36)$$

which only contains a kinetic term, with $F_{\alpha\beta} \equiv \partial_\alpha A_\beta - \partial_\beta A_\alpha$.

The part of the energy-momentum tensor that is proportional to the metric is already symmetric. The second term of (2.31) is given by

$$\Theta_{e.m.}^{\mu\alpha} = -\frac{\partial \mathcal{L}_{e.m.}}{\partial \partial_\mu A_\rho} \partial^\alpha A_\rho - \frac{1}{4} \eta^{\mu\alpha} F_{\alpha\beta} F^{\alpha\beta},$$

which can be expressed, after a simple rearrangement, as

$$\Theta_{e.m.}^{\mu\alpha} = F^{\alpha\rho} F_\rho^\mu - \frac{1}{4} \eta^{\mu\alpha} F_{\alpha\beta} F^{\alpha\beta} - F^{\rho\mu} \partial_\rho A^\alpha.$$

Here again, only the last term prevents the tensor from being symmetric (note, furthermore, that it is not even gauge invariant).

Let us now look for a tensor with three indices $G^{\rho\mu\alpha}$, antisymmetric in its first two indices, which will allow us to remove the last term. In other words, it should satisfy $\partial_\rho G^{\rho\mu\alpha} = F^{\rho\mu}\partial_\rho A^\alpha$. Using Maxwell's equations, this can be easily solved to obtain $G^{\rho\mu\alpha} = F^{\rho\mu}A^\alpha$. The symmetrized energy-momentum tensor is hence

$$T_{\text{e.m.}}^{\mu\nu} = F^{\nu\rho}F_\rho^\mu - \frac{1}{4}\eta^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}, \quad (2.37)$$

which turns out to be the same as the one obtained by variation with respect to the metric (1.83).

2.1.2 Quantum physics

2.1.2.1 Uncertainty principle and Hilbert spaces

The great conceptual revolution of quantum mechanics [3] arose from its statement that one cannot determine simultaneously with arbitrary precision the positions and momenta of any objects. This means, for instance, that if one measures the position and momentum of a particle with precisions Δq and Δp , then, at best (in one dimension),

$$\Delta q \cdot \Delta p \geq \frac{1}{2}\hbar, \quad (2.38)$$

where \hbar is Planck's constant. The existence of such an intrinsic limitation in the possibility of knowing the state of a physical system obliges us to give up the notion of a classical trajectory. (See, however, Ref. [4], which discusses the Bohm-de Broglie ontological interpretation for which such notions are still meaningful.)

The quantum description replaces the classical hypotheses, by assuming that any physical system is described by an element of a complex vector space, a space of states, i.e. configurations, the dimension of which depends on the system. Let us imagine, for instance, that we would like to study a light beam polarized linearly, either vertically or horizontally. If this is the only property we describe on the system, we then have a system with two possible states, denoted as $|\leftrightarrow\rangle$ and $|\uparrow\rangle$, in the 'kets' notation introduced by Dirac; the reverse notation allows us to define a scalar product, for example (\uparrow | represents a 'bra', so that the set of them forms a 'bra-c-ket'). The vector space, in which this system evolves, can have these two states as a basis, thus it is a two-dimensional space. Any element of this space can always be decomposed as $|\Psi\rangle = \alpha|\leftrightarrow\rangle + \beta|\uparrow\rangle$ with $\alpha, \beta \in \mathbb{C}$. The associated 'bra' is then obtained by the Hermitian conjugate, which is $\langle\Psi| = \alpha^*\langle\leftrightarrow| + \beta^*\langle\uparrow|$, so that the scalar product is given by the number $\langle\Psi|\Psi\rangle = |\alpha|^2\langle\leftrightarrow|\leftrightarrow\rangle + |\beta|^2\langle\uparrow|\uparrow\rangle$ for orthogonal basis states. We usually choose these basis vectors to be normalized to unity.

When describing the dynamics of the position or the momentum of a particle, each of these quantities corresponds to a 'coordinate' in the relevant vector space. This means that the basis, in which the system can be described, consists of all the elements $\{|q\rangle\}$. Since the coordinate q varies continuously in \mathbb{R} , the vector space has an

unaccountably infinite dimension. In this case (infinite dimension and endowed with a norm), it is called a Hilbert space.²

2.1.2.2 Operators and the superposition principle

In the quantum framework, the physical variables are replaced by operators that act on the vector space of states. For instance, for the state $|q\rangle$, which describes a particle at the position q , there exists a position operator, denoted as \hat{q} such that

$$\hat{q}|q\rangle = q|q\rangle. \quad (2.39)$$

In other words, the position state becomes an eigenstate of the position operator. To avoid any risk of confusion and error in what follows, we will systematically denote a quantum quantity corresponding to a classical one by the same letter with the symbol ‘~’: $q \mapsto \hat{q}$, $p \mapsto \hat{p}$, etc.

Any operator \hat{A} should satisfy some conditions in order to correspond to a measurable quantity. In particular, since the space of states is complex, the eigenvalues of an observable, which can potentially be measured, should be required to be real. This implies that \hat{A} is Hermitian, i.e. $\hat{A}^\dagger \equiv {}^T\hat{A}^* = \hat{A}$. Furthermore, since the eigenvalues α_k of this operator correspond to all possible values of the concerned property, the eigenstates $|\alpha_k\rangle$ of \hat{A} should form an orthonormal basis of the space of states. In other words, any arbitrary physical state $|\Psi\rangle$ can be decomposed as

$$|\Psi\rangle = \sum_k \psi_k |\alpha_k\rangle, \quad (2.40)$$

where the discrete sum should be replaced by an integral in the case of a continuous infinity of degrees of freedom i.e. in the case of a Hilbert space of states. An operator \hat{A} that satisfies these conditions is called an observable.

The relation (2.40) directly leads to the central property of quantum theory, namely the *superposition principle*, according to which several physical configurations can be superposed in a linear combination that is still a physical state; this is the case in (2.40). Moreover, a measurement of the observable \hat{A} performed on the state $|\Psi\rangle$ will necessarily give a unique eigenvalue α_k of this operator. Based on this principle, we can see that the dynamical equation satisfied by the physical state is necessarily linear.

2.1.2.3 Non-commutativity

Any given classical physical quantity corresponds to an operator in the space of states, and the equations that control the dynamics must reproduce those of classical physics in the limit where we know they are valid. The easiest way to obtain the quantum laws is thus to derive them from the corresponding classical laws. Now, unlike the functions that define classical physics, operators act on a vector space and do not always commute. It is this property that is used to realize the correspondence.

We define the commutator between two operators \hat{A} and \hat{B} as $[\hat{A}, \hat{B}] \equiv \hat{A}\hat{B} - \hat{B}\hat{A}$. This object has the same symmetry properties as the Poisson brackets introduced in

²The phrase ‘Hilbert space’ is also very often used for finite-dimensional spaces, although this is mathematically incorrect.

(2.5). It turns out that we obtain an appropriate prescription if we make the substitution

$$\{A, B\} \longmapsto -\frac{i}{\hbar} [\hat{A}, \hat{B}]. \quad (2.41)$$

In particular, the position and momentum variables q_i and p^i are now variables that no longer commute and (2.7) becomes

$$[\hat{q}_i, \hat{q}_j] = 0 = [\hat{p}^i, \hat{p}^j] \quad \text{and} \quad [\hat{q}_i, \hat{p}^j] = i\hbar\delta_i^j. \quad (2.42)$$

This relation can be generalized to numerous cases (see for example the case of a scalar field later on). One can explicitly check here that in the representation where the position operator is expressed by (2.39), the momentum operator then becomes $\hat{p}^i = -i\hbar\partial/\partial q_i$.

2.1.2.4 Predictability and statistics

The predictions of quantum mechanics are essentially probabilistic. Formally, this translates into saying that if a physical system is in a state $|\Psi\rangle$, then the probability $P_\Psi(\Phi)$ for it to be measured in a state $|\Phi\rangle$ is proportional to the square of the scalar product $\langle\Phi|\Psi\rangle$. Since only these probabilities are effectively predicted by the theory, one can choose to normalize these physical states, which will always be assumed in what follows, i.e. $\langle\Psi|\Psi\rangle = 1$. With this convention, we have

$$P_\Psi(\Phi) = |\langle\Phi|\Psi\rangle|^2. \quad (2.43)$$

Note that the normalization of a state is always possible, also because the scalar product of a state with itself is by definition positive in a vector space with positive-definite norm (which is a pre-requisite here to define probabilities)

$$\langle\Psi|\Psi\rangle = (|\Psi\rangle)^\dagger|\Psi\rangle = |||\Psi\rangle||^2 \geq 0.$$

For an observable operator \hat{A} , we have seen that the only values a measurement could give were those belonging to the spectrum (the set of possible eigenvalues) of this operator. Suppose that we have a great number of particles prepared in the same physical state, for definiteness, we choose, for instance, the state $|\Psi\rangle$ of the decomposition (2.40). The measurement of \hat{A} on this set of particles will be given by the average of the operator in the state $|\Psi\rangle$, namely

$$\langle\hat{A}\rangle_\Psi = \langle\Psi|\hat{A}|\Psi\rangle,$$

with a variance around this average value given by

$$(\Delta\hat{A})^2 \equiv \langle\Psi|(\hat{A} - \langle\hat{A}\rangle_\Psi)^2|\Psi\rangle = \langle\hat{A}^2\rangle_\Psi - \langle\hat{A}\rangle_\Psi^2.$$

As we would expect, this variance vanishes when $|\Psi\rangle$ is an eigenstate of \hat{A} .

The uncertainty relation for the observables \hat{A} and \hat{B} is obtained in a similar way [5] by considering the operator $\hat{C} = \Delta\hat{B}(\hat{A} - \langle\hat{A}\rangle) + i\Delta\hat{A}(\hat{B} - \langle\hat{B}\rangle)$, where from now

on (unless specified otherwise) the average is taken in an arbitrary state $|\Psi\rangle$, and the redundant index has been suppressed. It is sufficient to notice that $\langle \hat{C}^\dagger \hat{C} \rangle = \|\hat{C}|\Psi\rangle\|^2 \geq 0$ and to expand this inequality to obtain

$$\Delta \hat{A} \Delta \hat{B} \geq \frac{1}{2} \left| \langle i [\hat{A}, \hat{B}] \rangle \right|, \quad (2.44)$$

which, once applied to the position and momentum operators, reduces to the Heisenberg inequality (2.38), with an extra factor δ_i^j , arising from the fact that the coordinates and momentum in orthogonal directions commute.

2.1.2.5 Heisenberg and Schrödinger representations

Any physical state can always be normalized to unity. This allows for a probabilistic interpretation, which makes sense only if the dynamical evolution conserves the normalization of the states. Denoting by $\hat{U}(t)$ the time evolution operator, the fact that the evolution is linear implies that the state $|\Psi(t)\rangle$ evolves as

$$|\Psi(t)\rangle = \hat{U}(t - t_0)|\Psi(t_0)\rangle, \quad (2.45)$$

where \hat{U} satisfies

$$\langle \Psi(t) | \Psi(t) \rangle = \langle \Psi(t_0) | \Psi(t_0) \rangle \iff \hat{U}^\dagger \hat{U} = 1. \quad (2.46)$$

Thus the evolution operator \hat{U} is a unitary operator. In order to be physical, any quantum theory must satisfy this unitarity principle; this criterion allows us to rule out some theories, usually very speculative, without even needing to resort to experiments.

Unitary transformations actually allow us to go further and to define some representations i.e. some pairs $(\{\hat{A}_n\}, \{|\Psi\rangle_k\})$ that contain all possible operators and states of the theory. These representations are related to each other by

$$\begin{cases} |\Psi\rangle_k \longrightarrow |\tilde{\Psi}\rangle_k = \hat{U}|\Psi\rangle_k, \\ \hat{A}_n \longrightarrow \tilde{A}_n = \hat{U}\hat{A}_n\hat{U}^\dagger. \end{cases} \quad (2.47)$$

These relations guarantee the conservation of the norm of the states, and thus of the probabilities: all predictions made in a given representation will be identical to those obtained in another representation.

There exist many useful representations, in particular, the Heisenberg picture in which the operators are time dependent but where the basis vectors, the states, are time independent. In this representation, thanks to (2.41), the classical equation is directly transposed to

$$i\hbar \frac{d\hat{A}}{dt} = [\hat{A}, \hat{H}], \quad (2.48)$$

where \hat{H} is the operator that corresponds to the Hamiltonian function of the classical system. To obtain (2.48), we have assumed that the operator \hat{A} did not depend explicitly on time. Such an explicit dependence can be taken into account by simply adding the required term with a partial derivative.

The Schrödinger picture corresponds to the other extreme in the possibility of choices. In this representation the operators do not depend on time, and only the

states are time dependent. The difference between these two representations is similar to the choice one can make when describing rotations. Either one considers an active rotation of the system, leaving the axis unchanged, or one can describe the same rotations in a passive way, considering that it is the axes that change (in the opposite direction) while the system itself does not move.

2.1.2.6 Schrödinger equation

In the Schrödinger representation, the quantum dynamics of the state, now depending on time, simply follows from the requirement that predictions, and thus probabilities, should be time-translation invariant. Indeed, (2.45) implies that the time propagator $\hat{U}(t - t_0)$ satisfies

$$\frac{d}{dt} |\Psi(t)\rangle = \frac{d}{dt} \hat{U}(t - t_0) |\Psi(t_0)\rangle.$$

However, if the equation describing the dynamics of a state is linear and independent of the initial state and time t_0 , then the right-hand side term in the previous relation should be proportional to the vector $|\Psi(t)\rangle$. We can hence conclude that

$$\frac{d}{dt} \hat{U}(t - t_0) \propto \hat{U}(t - t_0).$$

This differential equation can be solved very easily to give

$$\hat{U}(t - t_0) = \exp \left[-\frac{i}{\hbar} \hat{H} \times (t - t_0) \right], \quad (2.49)$$

where \hat{H} is a Hermitian operator. From the time-translation invariance, this operator should be conserved, and should therefore be related to the energy. Actually it can only be the Hamiltonian. The form of (2.49) guarantees that the operator \hat{U} is indeed unitary.

Let us now consider again (2.45) with the solution (2.49), we obtain the Schrödinger equation

$i\hbar \frac{d}{dt} |\Psi(t)\rangle = \hat{H} |\Psi(t)\rangle.$

(2.50)

Moreover, we can note that \hat{U} is precisely the unitary operator one needs to shift from the Schrödinger representation to the Heisenberg one. The vector $|\Psi(t_0)\rangle$, which is effectively time independent, forms the basis of this representation.

Equation (2.50) can also be understood as a formal correspondence rule between the Hamilton function, i.e. the energy, and the operator $i\hbar d/dt$.

2.1.2.7 Harmonic oscillator

Before attacking the subjects of relativistic quantum mechanics and field theory, let us first recall the results of a very simple system, the harmonic oscillator. This will allow us to put into practice the notions discussed previously. Furthermore, this system will also be useful in field theory since all the fields, whose properties we shall use, can be expanded as sometimes infinite sets of such harmonic oscillators.

The Hamiltonian of a one-dimensional harmonic oscillator of mass m and angular frequency ω is

$$H = \frac{p^2}{2m} + \frac{1}{2}m\omega^2q^2. \quad (2.51)$$

The Schrödinger equation (2.50) for this Hamiltonian is obtained from the correspondences discussed previously, namely $H \rightarrow i\hbar d/dt$ and $p \rightarrow -i\hbar\partial/\partial q$,

$$i\hbar\frac{d\psi}{dt} = -\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial q^2} + \frac{1}{2}m\omega^2\psi, \quad (2.52)$$

where we have defined the wavefunction $\psi(q, t) \equiv \langle q|\psi(t)\rangle$. This function is usually complex. Given the interpretation (2.43) of quantum mechanics, the square of its modulus, $|\psi|^2$, represents the density of probability to find the particle at position q at time t . The normalization of the state $|\psi(t)\rangle$ simply implies that the total probability, integrated over the whole space, is unity, so that $\int |\psi|^2 dq = 1$.

The spectrum of the Hamiltonian, i.e. the set of the eigenvalues associated with the operator \hat{H} , is here obtained by solving the eigenvalue equation $\hat{H}\psi = E\psi$, or equivalently by looking for the oscillation modes of the wavefunction of the form $\psi = f(q)\exp(-iEt/\hbar)$. The equation we obtain is well known and its solutions are normalizable only if the energy is positive, which is physically acceptable. Furthermore, the derivative of the solution at $q = 0$ is continuous only for a discrete set of values of the energy, E_n .

We define the operators \hat{a} and \hat{a}^\dagger , called annihilation and creation operators, respectively, (understood as being associated with each oscillation mode), by the relations

$$\hat{a} \equiv \sqrt{\frac{m\omega}{2\hbar}} \left(\hat{q} + i\frac{\hat{p}}{m\omega} \right), \quad \hat{a}^\dagger = \sqrt{\frac{m\omega}{2\hbar}} \left(\hat{q} - i\frac{\hat{p}}{m\omega} \right); \quad (2.53)$$

they are conjugate to each other but not Hermitian: they are not observables. Inverting these relations and using the commutation rules (2.42), we find that $[\hat{a}, \hat{a}^\dagger] = 1$. The Hamiltonian can then be expressed as

$$\hat{H} = \hbar\omega \left(\hat{a}^\dagger \hat{a} + \frac{1}{2} \right).$$

In this form, we can understand why the energies are positive-definite since

$$\bar{E}_\psi = \langle \psi | \hat{H} | \psi \rangle = \hbar\omega \left(\langle \psi | \hat{a}^\dagger \hat{a} | \psi \rangle + \frac{1}{2} \right) = \hbar\omega \left(||\hat{a}|\psi\rangle||^2 + \frac{1}{2} \right) > 0,$$

by the definition of the norm in a Hilbert space.

Once we obtain a state $|n\rangle$ of energy E_n , we can determine all the others by applying the creation operator:

$$\hat{H}\hat{a}^\dagger|n\rangle = \hbar\omega \left(\hat{a}^\dagger \hat{a} + \frac{1}{2} \right) \hat{a}^\dagger|n\rangle = (E_n + \hbar\omega) \hat{a}^\dagger|n\rangle,$$

which shows that $|n+1\rangle = (n+1)^{-1/2}\hat{a}^\dagger|n\rangle$ is an eigenstate of the Hamiltonian with the eigenvalue $E_{n+1} = E_n + \hbar\omega$. Similarly, $|n-1\rangle = n^{-1/2}\hat{a}|n\rangle$ is the eigenstate of \hat{H}

with the energy $E_{n-1} = E_n - \hbar\omega$. The state with lowest energy, the vacuum state $|0\rangle$, satisfies $\hat{a}|0\rangle = 0$, and all the states can be obtained by

$$|n\rangle = \frac{1}{\sqrt{n!}} (\hat{a}^\dagger)^n |0\rangle.$$

Thus, the vacuum state has a non-vanishing energy, $E_0 = \frac{1}{2}\hbar\omega$, and the full spectrum is $E_n = (n + \frac{1}{2})\hbar\omega$.

2.1.2.8 General potential

Let us now come back to the three-dimensional space. A particle evolving in a general potential $V(x)$ can be expanded in a similar way in the basis of the eigenstates of the Hamiltonian, with eigenfunctions, $\psi_E(x^\mu)$ satisfying

$$\left[-\frac{\hbar^2}{2m} \nabla^2 + V(x) \right] \psi_E(x^\mu) = E \psi_E(x^\mu), \quad (2.54)$$

where the eigenvalue of the Hamiltonian E , is the energy of the state.

For a harmonic oscillator, $V_{\text{h.o.}} \propto x^2$, and (2.54) becomes invariant under the transformations $x \rightarrow -x$ and $p \rightarrow p$. Therefore, the eigenstates then have vanishing average values of the position and momentum, i.e. $\langle n|\hat{q}|n\rangle = 0 = \langle n|\hat{p}|n\rangle$. This reflects the fact that the potential has a minimum at the origin, increases with the distance (quantum analogue of a restoring force) and is symmetric. Thus, the wavefunctions are essentially localized at the origin and the requirement that they are normalizable demands that they vanish asymptotically.

A general potential $V(x)$ does not necessarily respect these conditions. In particular, it can be non-vanishing only in a finite region of space. This is the case, for instance, for the Coulomb potential of the electric interaction between an electron and a proton. This attractive potential is, in addition, negative. We therefore deduce that the eigenvalues are negative energies forming the different possible states of the hydrogen atom, spread on a discrete spectrum of normalizable wavefunctions. There is also a continuous component formed by all the positive energies, which corresponds to non-normalizable wavefunctions. Actually, they are the wavefunctions of the momentum operator that can be expressed as

$$\psi_p(x^\mu) \propto e^{i(p \cdot x - \omega t)},$$

i.e. plane waves. We usually normalize them by artificially putting the particle in a box of finite volume V_3 , and dividing the wavefunction by V_3 . Once all computations are performed nothing should depend any longer on V_3 and one can take the limit $V_3 \rightarrow \infty$. If this is not the case, using the wavefunctions of the Hamiltonian is delicate and one should resort to other methods and different representations.

An important point, valid both in classical and in quantum mechanics, is that the potential $V(x)$ must be determined by experiments, or should be imposed in one way or another. Its functional form is in the theoretical context completely arbitrary. The description of some interactions in the framework of field theory will allow us to compute the form of this potential for many cases.

The eigenstates of the momentum can also be used as a basis for the representation of free particles, i.e. without a potential. In this case, the creation operator $\hat{a}^\dagger(p)$ acts on the vacuum state to create a particle with momentum p , and the wavefunction is obtained using the Fourier transform of these states. It follows that

$$\psi(x, t) = \langle x | \psi(t) \rangle = \int \frac{d^3 p}{(2\pi)^{3/2}} e^{ip \cdot x} \langle p | \psi(t) \rangle = \int \frac{d^3 p}{(2\pi)^{3/2}} e^{ip \cdot x} \langle 0 | \hat{a}_p | \psi(t) \rangle, \quad (2.55)$$

which allows us to define the field operator $\hat{\Psi}$ by

$$\hat{\Psi}(x) \equiv \int \frac{d^3 p}{(2\pi)^{3/2}} e^{ip \cdot x} \hat{a}_p, \quad (2.56)$$

such that $\psi(x, t) = \langle 0 | \hat{\Psi}(x) | \psi(t) \rangle$. The operator $\hat{\Psi}(x)$ annihilates a particle at the point x , while its Hermitian conjugate $\hat{\Psi}^\dagger$ produces a particle at the same point.

The generalized commutation relations for the creation and annihilation operators for momentum eigenstates, namely

$[\hat{a}_p, \hat{a}_k] = 0 = [\hat{a}_p^\dagger, \hat{a}_k^\dagger], \quad [\hat{a}_p, \hat{a}_k^\dagger] = \delta^{(3)}(p - k),$

(2.57)

translate into

$$[\hat{\Psi}(x), \hat{\Psi}(y)] = 0 = [\hat{\Psi}^\dagger(x), \hat{\Psi}^\dagger(y)], \quad [\hat{\Psi}(x), \hat{\Psi}^\dagger(y)] = \delta^{(3)}(x - y). \quad (2.58)$$

Note that it is also possible to define a conjugate momentum $\hat{P}(x)$ and a Hamiltonian formalism. We will come back to this in the context of the relativistic theory.

2.1.3 Quantum and relativistic mechanics

The application of the correspondence principle allows the immediate formulation of a relativistic quantum theory. Instead of using the classical expressions relating the total energy, the Hamiltonian, to the kinetic and potential energies, it is sufficient to consider the relativistic momentum, $p^\mu = mu^\mu$, and the invariant $p_\mu p^\mu = -m^2 c^2$, which can be expressed as a function of the energy, E , and of the momentum, p , as $p_\mu p^\mu = -E^2/c^2 + p^2 = -m^2 c^2$, to get

$$E^2 = p^2 c^2 + m^2 c^4.$$

The Klein–Gordon equation for the wave function ϕ can then be obtained using the correspondence $E \rightarrow i\hbar \partial_t$ and $p \rightarrow -i\hbar \nabla$

$$\left(-\frac{\hbar^2}{c^2} \frac{\partial^2}{\partial t^2} + \hbar^2 \nabla^2 - m^2 c^2 \right) \phi = 0. \quad (2.59)$$

It will be shown in what follows how it is possible to take into account the properties of special relativity to describe quantum field theory. An important part of these developments can be generalized to curved space-times, in particular by replacing the partial derivatives ∂_α by covariant ones ∇_α . This is how we can consider the influence

of gravity by means of general relativity to a large degree: the quantum field theory in curved space-time assumes that gravitation alters field theory only through the metric in which the (test) quantum fields evolve.

From now on, natural units will be considered and we therefore fix ‘ $\hbar = c = 1$ ’ (see appendix A).

2.1.3.1 Klein-Gordon and Dirac equations

In natural units, the Klein-Gordon equation, which describes the dynamics of a free scalar field, becomes

$$\left(-\frac{\partial^2}{\partial t^2} + \nabla^2 - m^2 \right) \phi = (\partial_\mu \partial^\mu - m^2) \phi \equiv (\square - m^2) \phi = 0. \quad (2.60)$$

For a general self-interacting potential, the Lagrangian from which this equation is derived is given by (1.106). In the case of a massive free scalar field, we will take

$$\mathcal{L}_{\text{KG}} = -\frac{1}{2} \partial_\mu \phi \partial^\mu \phi - \frac{1}{2} m^2 \phi^2, \quad (2.61)$$

i.e. a simple mass term. Its quantization will be discussed in detail in Section 2.2.

A problem immediately appears in (2.60), that of the existence of negative-energy solutions. Indeed, the classical relation $E = \sqrt{p^2 + m^2}$ assumes that $E > 0$. But the solutions of (2.60) are simply the eigenstates of the momentum and the energy, which are $\propto \exp[i(p \cdot x - Et)]$, but nothing indicates what the sign of the energy should be. In other words, both choices $E = \pm \sqrt{p^2 + m^2}$ are possible. But if the wavefunction ϕ corresponds to one of these states, then it contains less energy than the vacuum state itself. Thus, a theory containing such states is necessarily unstable. This is why we should go to the level of quantum field theory, for which the solution of (2.60) is not a wavefunction, as in the case of ψ for the Schrödinger equation, but an operator, similar to $\hat{\Psi}$ of (2.56).

The difficulty we have just mentioned for the scalar field comes from the fact that its dynamical equation, unlike in the classical case, is second order in time, which is normal to respect the relativistic invariance. Another way to proceed is to postulate a first-order equation in time, and hence also in space. In order to respect the relativistic invariance, one should be able to write it as

$$(i\gamma_\mu \partial^\mu + m) \psi = 0, \quad (2.62)$$

which we can obtain by varying the Dirac Lagrangian

$$\mathcal{L}_{\text{Dirac}} = -\bar{\psi} (i\gamma^\mu \partial_\mu + m) \psi. \quad (2.63)$$

Since this equation, known as the Dirac equation, describes the dynamics of a relativistic particle its square should also reproduce the correspondence observed in the Klein-Gordon equation. We should therefore also have $(\eta^{\mu\nu} \partial_\mu \partial_\nu - m^2)\psi = 0$, which implies that

$$(i\gamma^\mu \partial_\mu - m)(i\gamma^\nu \partial_\nu + m)\psi = -(\gamma^\mu \gamma^\nu \partial_\mu \partial_\nu + m^2)\psi = 0.$$

This is only compatible with the Klein–Gordon equation if the coefficients γ^μ satisfy

$$\{\gamma^\mu, \gamma^\nu\} \equiv \gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu = -2\eta^{\mu\nu}. \quad (2.64)$$

Thus, the γ_μ cannot be simple numbers. Actually, they should be matrices of rank at least equal to four. A useful representation for these γ^μ is

$$\gamma^0 = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}, \quad \gamma^i = \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix} \implies \gamma^\mu = \begin{pmatrix} 0 & \sigma^\mu \\ \bar{\sigma}^\mu & 0 \end{pmatrix}, \quad (2.65)$$

where each entry of these matrices is itself a 2×2 matrix, $I = \sigma^0$ standing for the two-dimensional identity matrix, i.e. $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. The σ^i are the Pauli matrices, that is

$$\sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \text{and} \quad \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (2.66)$$

Finally, we have defined the vectors $\sigma^\mu \equiv (I, \sigma^i)$ and $\bar{\sigma}^\mu \equiv (I, -\sigma^i)$. For reasons that will become clear later, the matrices of (2.65) are called the *chiral representation* of the Clifford algebra, defined by (2.64).

It is interesting to note that (2.62) can also be obtained from the action

$$\mathcal{S}_D = - \int d^4x \left(\frac{1}{2} \bar{\psi} i \gamma_\mu \overleftrightarrow{\partial}^\mu \psi + m \bar{\psi} \psi \right). \quad (2.67)$$

Assuming that $\bar{\psi}$, defined by

$$\bar{\psi} \equiv \psi^\dagger \gamma^0, \quad (2.68)$$

and ψ are considered as two independent quantities, then the variations must be performed simultaneously. This action introduces the operator

$$F \overleftrightarrow{\partial} G \equiv F \partial(G) - \partial(F)G, \quad (2.69)$$

for any two quantities F and G . In the case of (2.67), F and G are spinors, but this definition applies also for any kind of field, for instance scalars.

2.1.3.2 Maxwell's and Proca's theories

The scalar field ϕ is of spin 0 and the spinor field ψ of spin 1/2. The fields known experimentally, among which we can find all the particles listed in Section 2.3, also include particles of spin 1 such as the photon, of zero mass, or the W^\pm and Z^0 vectors of the electroweak interaction. Such fields are described by a vector A^μ , the dynamics of which is governed by Proca's Lagrangian

$$\mathcal{L}_{\text{Proca}} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - \frac{1}{2} M_A^2 A_\mu A^\mu \quad (2.70)$$

for a real field of mass M_A . The Faraday tensor has the usual definition, $F_{\mu\nu} \equiv 2\partial_{[\mu} A_{\nu]}$.

Varying (2.70) using (2.14) for each component A_α of the vector field, we find

$$\frac{\partial}{\partial(\partial^\mu A^\alpha)} \left(-\frac{1}{4} F_{\rho\sigma} F^{\rho\sigma} \right) = -F_{\mu\alpha},$$

and hence

$$(\square - M_A^2) A_\alpha = \partial^\mu \partial_\alpha A_\mu. \quad (2.71)$$

This equation does not seem to be equivalent to the Klein–Gordon equation for each component of the field due to the right-hand side term. However, if we take the divergence ∂^α of this equation, the right-hand side term becomes a d'Alembertian, which simplifies with the one on the left, so that

$$M_A^2 \partial^\alpha A_\alpha = 0.$$

Note that things are more complicated in the relativistic case, where the covariant derivatives do not commute [cf. (1.97)]. As can be seen here, if the vector field is massive, i.e. $M_A \neq 0$, then the Lorentz gauge condition $\partial^\mu A_\mu = 0$ is automatically satisfied [the fact that Proca's theory (2.70) is not gauge invariant precisely because of this term is no surprise], and (2.71) leads to a Klein–Gordon equation for each component, which is to be expected from the classical-quantum correspondence principle discussed around (2.59).

The energy-momentum tensor for the massive vector field, once it has been suitably symmetrized, is identical to the one for the gauge field of electromagnetism (2.37) up to the mass term that has to be added. This leads to

$$T_{\text{Proca}}^{\mu\alpha} = F^{\alpha\rho} F_\rho^\mu - \frac{1}{4} \eta^{\mu\alpha} F_{\alpha\beta} F^{\alpha\beta} - M_A^2 \left(A^\mu A^\alpha - \frac{1}{2} \eta^{\mu\alpha} A_\nu A^\nu \right). \quad (2.72)$$

2.2 Canonical decompositions

The particular case of a scalar field, let it be real or complex, is discussed in detail in most textbooks on field theory; see, for instance, Refs. [6–9]. We briefly review it in this section. For completeness, we also briefly describe the case of the vector and spinor fields.

2.2.1 Technical clarification

In what follows, we will systematically choose a symmetric definition of the Fourier transform and of its inverse that are defined by (B.58) and (B.59).

2.2.1.1 Lorentz invariant measure

One of the important properties of the Dirac distribution is that

$$\delta[f(x) - \alpha] = \sum_a \left(\frac{df}{dx} \right)^{-1}_{x=x_a} \delta(x - x_a), \quad (2.73)$$

where the x_a represent the zeros of the argument, i.e. the solutions of $f(x_a) = \alpha$.

This property allows us to replace the Fourier integrals in 4 dimensions by three-dimensional integrals when the fields over which we perform the integration are ‘on shell’, i.e. their Fourier components satisfy the relation $E^2 \equiv k_0^2 = \mathbf{k}^2 - m^2$, the energy being positive-definite. In that case, it is possible to perform the integration over the time component k_0 with the previous relation by setting $f(k_0) = k_0^2$. This gives

$$\begin{aligned} \int \frac{d^4k}{(2\pi)^{3/2}} \delta(k^2 - m^2) \theta(k_0) &= \int \frac{d^3k}{(2\pi)^{3/2}} dk_0 \delta[k_0^2 - (k^2 - m^2)] \theta(k_0) \\ &= \int \frac{d^3k}{(2\pi)^{3/2}} \frac{1}{2\omega_k}, \end{aligned} \quad (2.74)$$

with the definition

$$\omega_k \equiv +\sqrt{k^2 - m^2}, \quad (2.75)$$

and where θ is the Heaviside step function. The relation (2.74) defines a Lorentz invariant measure despite being three-dimensional. Note that to obtain this measure, it is necessary to impose the positive energy condition in order to keep only one term of (2.73).

2.2.2 Real free field

Equation (2.60) can be obtained by varying the Lagrangian (1.106) in which we postulate an arbitrary potential $V(\phi)$ for the scalar field ϕ . The conjugate field is then

$$\pi(x) = \frac{\delta \mathcal{L}}{\delta \dot{\phi}(x)} = \dot{\phi}(x), \quad (2.76)$$

so that the Hamiltonian density, $\mathcal{H} = \pi \dot{\phi} - \mathcal{L}$, is given by

$$\mathcal{H} = \frac{1}{2} \left[\pi^2 + (\nabla \phi)^2 + 2V(\phi) \right]. \quad (2.77)$$

General potentials $V(\phi)$ are useful in cosmology, for instance, when addressing the inflation scenarios (Chapter 8).

2.2.2.1 Mode expansion

The eigenstates of the momentum operator are the basis of every expansion in quantum field theory. To deduce that, we should first know the energy-momentum tensor, that is

$$\begin{aligned} \Theta^{\mu\alpha} &= -\frac{\partial \mathcal{L}}{\partial \partial_\mu \phi} \partial^\alpha \phi + \eta^{\mu\alpha} \mathcal{L}, \\ &= \partial^\mu \phi \partial^\alpha \phi - \eta^{\mu\alpha} \left[\frac{1}{2} (\partial \phi)^2 + V(\phi) \right] = T^{\mu\alpha}. \end{aligned} \quad (2.78)$$

The last equality follows from the fact that the energy-momentum tensor of a scalar field is already symmetric. Moreover, it coincides with the tensor (1.107) obtained by varying (1.84) with respect to the metric.

The tensor (2.78) allows us to construct the four-momentum vector that is obtained by $P^\alpha \equiv T^{0\alpha}$, from which we indeed get, first the Hamiltonian density, which corresponds to the purely time-like component, i.e. $P^0 \equiv T^{00} = \mathcal{H}$, and then the associated momentum to the scalar field $\mathbf{P} = \pi \nabla \phi$.

Because the presence of a general potential $V(\phi)$ can lead to some potential nonlinearities that can imply some unnecessary complications, here and in the following sections we restrict ourselves to the case of the free massive field for which $V = \frac{1}{2}m^2\phi^2$. Having found the momentum \mathbf{P} , it is sufficient to perform the substitution of the correspondence principle and to replace \mathbf{P} by $\mathbf{P} \rightarrow -i\nabla$. The eigenmodes $u_k(x)$ with eigenvalue k of the momentum operator thus satisfy $-i\nabla u_k(x) = ku_k(x)$. They are plane waves

$$u_k(x) = N_k e^{ik \cdot x},$$

N_k being a normalization factor to be determined later. It is now sufficient to expand the field on the basis of these eigenstates as

$$\phi(x, t) = \int d^3k \alpha_k(t) u_k(x). \quad (2.79)$$

Since ϕ satisfies the Klein–Gordon equation (2.60), we find that the coefficients $\alpha_k(t)$ evolve as

$$\ddot{\alpha}_k(t) + (k^2 + m^2)\alpha_k(t) = 0,$$

where we used the fact that $\nabla^2 u_k(x) = -k^2 u_k(x)$. Setting $\omega_k^2 = k^2 + m^2$, the solution of this equation is simply

$$\alpha_k = a_k e^{i\omega_k t} + b_k e^{-i\omega_k t},$$

with $b_k^* = a_{-k}$ since $\phi \in \mathbb{R}$. After some algebraic manipulations, and using the relations between the coefficients of the expansion, we find a more explicit four-dimensional expression, that is

$$\phi(x) = \phi(x, t) = \int d^3k N_k (a_k e^{ik \cdot x} + a_k^* e^{-ik \cdot x}), \quad (2.80)$$

where we recall that $k \cdot x \equiv \mathbf{k} \cdot \mathbf{x} - \omega_k t$. From (2.80) and from the invariant measure (2.74) we see that the field expansion, even though it was realized on the basis of three-dimensional states, is invariant, i.e. that ϕ is effectively a Lorentz scalar, as long as we choose the normalization

$$N_k = \frac{1}{(2\pi)^{3/2} \sqrt{2\omega_k}}. \quad (2.81)$$

Moreover, this relation allows us to recover the usual commutation relations after quantization. Note that the conjugate momentum of the field is simply

$$\pi(x, t) = \int d^3k N_k (-i\omega_k) (a_k e^{ik \cdot x} - a_k^* e^{-ik \cdot x}), \quad (2.82)$$

which we obtain in a similar manner as before. The factor $(-\iota\omega_k)$ in the integrand of (2.82) implies that the latter is not explicitly a Lorentz scalar. This is not a problem since only the field should absolutely satisfy this property.

2.2.2.2 Quantization

The field and its conjugate momentum can now be ‘promoted’ to the status of operators, which can be done by replacing the coefficients a_k and a_k^* by the annihilation and creation operators, respectively, \hat{a}_k and \hat{a}_k^\dagger . The field Poisson brackets then become the equal time commutations relations. If we impose

$$[\hat{a}_k, \hat{a}_{k'}^\dagger] = \delta^{(3)}(k - k') , \quad (2.83)$$

and all others to vanish, i.e. the usual commutation relations, we indeed obtain the canonical relations for the field

$$[\hat{\phi}(x, t), \dot{\hat{\phi}}(x', t)] = i\delta^{(3)}(x - x') , \quad (2.84)$$

a computation that is recommended to be performed explicitly. The relations (2.83) and (2.84) can only be respected if the normalization of the field modes is given by (2.81). The eigenmodes of the field operator are now defined by

$$u_k(x, t) = N_k e^{i(k \cdot x - \omega_k t)} , \quad (2.85)$$

so that the field operator is expressed as

$$\hat{\phi}(x, t) = \int d^3 k [\hat{a}_k u_k(x, t) + \hat{a}_k^\dagger u_k^*(x, t)] . \quad (2.86)$$

This relation can be inverted to express the creation and annihilation operators as a function of the field, namely

$$\hat{a}_k = i \int d^3 x [u_k^*(x, t) \overleftrightarrow{\partial}_t \hat{\phi}(x, t)] , \quad (2.87)$$

where the derivative $\overleftrightarrow{\partial}$ is defined by (2.69).

This analysis is then complete as soon as we know the Hamiltonian, which is given by

$$\hat{H}(t) = \int d^3 x \hat{\mathcal{H}}(x, t) = \frac{1}{2} \int d^3 x \left[\dot{\hat{\phi}}^2 + (\nabla \hat{\phi})^2 + m^2 \hat{\phi}^2 \right] . \quad (2.88)$$

We then get

$$\hat{H}(t) = \frac{1}{2} \int d^3 k \omega_k (\hat{a}_k^\dagger \hat{a}_k + \hat{a}_k \hat{a}_k^\dagger) = \int d^3 k \omega_k \left[\hat{\mathcal{N}}_k + \frac{1}{2} \delta^{(3)}(0) \right] , \quad (2.89)$$

where we have used the commutation relation (2.83) and defined $\hat{\mathcal{N}}_k$ as the ‘number of particles in the state with momentum k ’ operator, the significance of which will lead us to the definition of the Fock space associated with this theory.

A remark concerning (2.89) is necessary at this point. The second term, proportional to $\delta^{(3)}(0)$, is clearly divergent, even in the case where the momentum space is finite, i.e. if there exists a cutoff k_{\max} above which one can no longer perform the computation. In field theory, as discussed here, this infinite energy is not embarrassing as

it has no physical meaning since in practice only differences of energy are measurable. In practice, we define a normal ordering that automatically puts the creation operators on the left and the annihilation ones on the right, or by ‘renormalizing’ the energy by simply suppressing the vacuum energy obtained this way. This possibility is unfortunately no longer an option as soon as the theory is coupled to gravity. The vacuum energy then assumes all its meaning and acts as a cosmological constant, which is in any case considerably more significant than the one we measure (Chapter 12). *It is, in the current state of knowledge, the most serious problem that theoretical cosmology has to confront, and it comes from quantum field theory!*

2.2.2.3 Fock space

The complete quantum theory is obtained once we know the state of the particles. For that, we start by defining the vacuum $|0\rangle$ as being the state annihilated by all the annihilation operators, i.e. $\hat{a}_k|0\rangle = 0 \forall k \in \mathbb{R}^3$. We consider, furthermore, that this state is normalized, namely $\langle 0|0\rangle = 1$. A state with one particle is then constructed from this vacuum state by applying a creation operator to it. We thus have

$$|k\rangle \equiv \hat{a}_k^\dagger|0\rangle, \quad (2.90)$$

normalized by

$$\langle k|k' \rangle = \langle 0|\hat{a}_k\hat{a}_{k'}^\dagger|0\rangle = \delta^{(3)}(k - k'),$$

and we construct in this way a complete space of states, each state having $n(k)$ particles in the state of momentum k . The definition (2.90) shows that a state with $n(k)$ is an eigenstate of the operator \hat{N}_k , since

$$\hat{N}_k|n(k)\rangle \equiv \hat{a}_k^\dagger\hat{a}_k|n(k)\rangle = n(k)|n(k)\rangle. \quad (2.91)$$

It follows that any state describing a set of particles will be an element of

$$\mathcal{F} \equiv \{|n_1(k_1), n_2(k_2), \dots\rangle\},$$

called a *Fock space* and having an unaccountable infinity of states.

Finally, note that starting from the vacuum state $|0\rangle$, it is possible to produce a particle at the point x simply by applying the field operator itself

$$|x, t\rangle = \phi^\dagger(x, t)|0\rangle = \int \frac{d^3k}{(2\pi)^{3/2}} \frac{e^{-ik \cdot x}}{\sqrt{2\omega_k}} |k\rangle, \quad (2.92)$$

which appears essentially as the Fourier transform of the one particle state with given momentum, hence its interpretation. Note that this equation uses the operator ϕ^\dagger to generate a particle, which plays no important role here since the field is real and hence $\phi^\dagger = \phi$, but in the more general case of a complex field, the distinction is important.

2.2.2.4 Spherical waves expansion

The previous expansion is only a special case of the expansion in the basis of the momentum eigenstates expressed in Cartesian coordinates (x, y, z) . It is also possible

to perform this expansion in spherical coordinates (r, θ, φ) using spherical harmonics. In that case, we split the Laplacian into radial and angular parts $\Delta = \Delta_r + r^{-2}\Delta_\Omega$, with $\Delta_r = \partial_r^2 + (2/r)\partial_r$ and $\Delta_\Omega = \partial_\theta^2 + \sin^2\theta\partial_\varphi^2$. We express the eigenvalues of the Laplacian using the amplitude $p \equiv \sqrt{p^2}$ and two integer numbers $\ell \in \mathbb{N}$ and $m \in [-\ell, +\ell]$. The scalar field $\phi(x, t)$ can then be decomposed into the eigenmodes, $\phi_{p\ell m}(x)$, of the Laplacian

$$(\Delta + p^2) \phi_{p\ell m}(x) = 0, \quad (2.93)$$

which are given by

$$\phi_{p\ell m}(x, t) = u_{p\ell}(r)Y_{\ell m}(\theta, \varphi).$$

Since the spherical harmonics are solutions of (B.8), (2.93) leads to an ordinary differential equation for the radial part $u_{p\ell}(r)$, namely

$$\left[\frac{d^2}{dr^2} + \frac{2}{r} \frac{d}{dr} - \frac{\ell(\ell+1)}{r^2} + p^2 \right] u_{p\ell} = 0, \quad (2.94)$$

the solution of which is a spherical Bessel function $u_{p\ell}(r) \propto j_\ell(pr)$, (Appendix B). The normalization relations of the spherical harmonics (B.15) together with those of the spherical Bessel functions (B.51) show that choosing

$$u_{p\ell}(r) = \sqrt{\frac{2}{\pi}} p j_\ell(pr),$$

the modes of the scalar field are normalized, i.e.

$$\int d^3x \phi_{p\ell m}^*(x) \phi_{p'\ell' m'}(x) = \delta(p - p') \delta_{\ell\ell'} \delta_{mm'}. \quad (2.95)$$

Also note that with these conventions, we have the completeness relation

$$\int_0^\infty dp \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{+\ell} u_{p\ell}(r) u_{p\ell}(r') Y_{\ell m}^*(\theta, \varphi) Y_{\ell m}(\theta', \varphi') = \delta^{(3)}(x - x'). \quad (2.96)$$

The scalar field operator is now expressed as

$$\hat{\phi}(x, t) = \int_0^\infty dp \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{+\ell} N_p u_{p\ell}(r) Y_{\ell m}(\theta, \varphi) \hat{a}_{p\ell m}(t), \quad (2.97)$$

where the operators $\hat{a}_{p\ell m}$ depend explicitly on time. Inserting the relation (2.97) in the Klein–Gordon equation (2.60), we get

$$\left(\frac{d^2}{dt^2} + \omega_p^2 \right) \hat{a}_{p\ell m}(t) = 0, \quad (2.98)$$

where $\omega_p^2 = p^2 + m^2$ for each independent mode. The relation (2.98) shows that the creation and annihilation operators, which could a priori depend on the three numbers p , ℓ and m , have a time dependence that is only a function of the magnitude p of the

momentum. The solutions of equation (2.98) are complex exponentials, so that the total field takes the form

$$\hat{\phi}(x, t) = \int_0^\infty dp \sum_{\ell, m} N_p u_{p\ell}(r) \left[Y_{\ell m}(\theta, \varphi) \hat{a}_{p\ell m} e^{-i\omega_p t} + Y_{\ell m}^*(\theta, \varphi) \hat{a}_{p\ell m}^\dagger e^{i\omega_p t} \right], \quad (2.99)$$

and the momentum conjugate

$$\begin{aligned} \hat{\pi}(x, t) &= \int_0^\infty dp \sum_{\ell, m} (-i\omega_p) N_p u_{p\ell}(r) \\ &\times \left[Y_{\ell m}(\theta, \varphi) \hat{a}_{p\ell m} e^{-i\omega_p t} - Y_{\ell m}^*(\theta, \varphi) \hat{a}_{p\ell m}^\dagger e^{i\omega_p t} \right]. \end{aligned} \quad (2.100)$$

The remaining thing to do is to use the commutation relations (2.84) at equal time and to impose the same normalization as previously, namely $N_p = (2\omega_p)^{-1/2}$, to get the commutation relations for the creation and annihilation operators

$$[\hat{a}_{p\ell m}, \hat{a}_{p' \ell' m'}^\dagger] = \delta_{\ell\ell'} \delta_{mm'} \delta(p - p'), \quad (2.101)$$

$$[\hat{a}_{p\ell m}, \hat{a}_{p' \ell' m'}] = [\hat{a}_{p\ell m}^\dagger, \hat{a}_{p' \ell' m'}^\dagger] = 0.$$

The link between the Cartesian and spherical representations is obtained by expanding the plane wave in the basis of spherical waves (B.21). By identifying the coefficients of the term $e^{-i\omega_p t}$ in both expansions (2.86) and (2.99), we obtain

$$\hat{a}_{p\ell m} = i^\ell p \int d\theta_p \sin \theta_p d\varphi_p Y_{\ell m}^*(\theta_p, \varphi_p) \hat{a}_p,$$

(2.102)

where (θ_p, φ_p) represent the angles of the vector p . Equation (2.102) gives the relation we were looking for between the creation and annihilation operators in both representations, and hence between the respective Fock spaces.

2.2.3 The complex scalar field

The generalization to the case of a complex scalar field, $\phi \in \mathbb{C}$, is automatic as soon as we decompose this field into its real and imaginary parts, i.e. as

$$\phi = \text{Re}(\phi) + i\text{Im}(\phi) \equiv \frac{1}{\sqrt{2}}(\phi_1 + i\phi_2),$$

and we assume that the real and imaginary parts are independent. In a perfectly equivalent way, and this is the point of view we will adopt, we can also consider ϕ and ϕ^* as two independent fields. The Lagrangian obtained in both cases should be real, so that for a free massive field

$$\mathcal{L} = -\partial_\mu \phi^* \partial^\mu \phi - m^2 |\phi|^2. \quad (2.103)$$

Note the absence of the factor $\frac{1}{2}$ in the mass term, compared with the Lagrangian (2.61). This is due to the fact that since the two degrees of freedom of the field are

independent, the variation with respect to ϕ or ϕ^* does not give the factor 2 of the real scalar field.

The two conjugate momentums are

$$\pi = \frac{\delta \mathcal{L}}{\delta \dot{\phi}} = \dot{\phi}^* \quad \text{and} \quad \pi^* = \frac{\delta \mathcal{L}}{\delta \dot{\phi}^*} = \dot{\phi},$$

which are indeed complex conjugates of one another and are hence independent. The theory describing a complex scalar field is simply that of two real scalar fields of the same mass, related to each other by the fact that they are the real and imaginary parts of a complex number.

The energy-momentum tensor is obtained by summing (2.31) over the independent degrees of freedom

$$\Theta_{\mu\nu}^{(c)} = -\frac{\partial \mathcal{L}}{\partial \partial^\mu \phi} \partial_\nu \phi - \frac{\partial \mathcal{L}}{\partial \partial^\mu \phi^*} \partial_\nu \phi^* + \eta_{\mu\nu} \mathcal{L},$$

which gives

$$\Theta_{\mu\nu}^{(c)} = \partial_\mu \phi^* \partial_\nu \phi + \partial_\mu \phi \partial_\nu \phi^* + \eta_{\mu\nu} \mathcal{L} = T_{\mu\nu}. \quad (2.104)$$

We find again the same result as the one obtained using the definition that comes from the variation with respect to the metric.

The commutation relations at equal time between the field operators and their conjugate momentum are now

$$[\hat{\phi}(x, t), \hat{\pi}(x', t)] = [\hat{\phi}^\dagger(x, t), \hat{\pi}^\dagger(x', t)] = i\delta^{(3)}(x - x'). \quad (2.105)$$

The mode expansion is of the form

$$\hat{\phi}(x, t) = \int \frac{d^3 k}{(2\pi)^{3/2} \sqrt{2\omega_k}} (\hat{a}_k e^{ik \cdot x} + \hat{b}_k^\dagger e^{-ik \cdot x}), \quad (2.106)$$

where we have introduced two different sets of operators, \hat{a}_k and \hat{b}_k , since the field is complex. Their commutation relations are now

$$[\hat{a}_k, \hat{a}_{k'}^\dagger] = [\hat{b}_k, \hat{b}_{k'}^\dagger] = \delta^{(3)}(k - k'), \quad (2.107)$$

all the other commutators vanish.

With the normal ordering for the operators discussed in the case of a real scalar field, the Hamiltonian takes a form equivalent to (2.89), that is

$$\hat{H}_c = \int d^3 k \omega_k (\hat{a}_k^\dagger \hat{a}_k + \hat{b}_k^\dagger \hat{b}_k) \equiv \int d^3 k \omega_k [\hat{N}_k^{(a)} + \hat{N}_k^{(b)}], \quad (2.108)$$

which shows that a complex scalar field can be interpreted as a collection of two kinds of different particles, but with the same mass, corresponding to the two sets of operators \hat{a}_k and \hat{b}_k . The Fock space is then the tensor product of the two Fock spaces generated by these operators, with one mutual state, which is the vacuum $|0\rangle$, annihilated by both \hat{a}_k and \hat{b}_k .

2.2.3.1 Maxwell or Proca field

The Lagrangian of a massive complex vector field is obtained by generalizing Proca's Lagrangian (2.71) as

$$\mathcal{L}_{\text{Proca}} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + M_A^2 A_\mu^* A^\mu, \quad (2.109)$$

where A_μ now has four complex components. As for the scalar field, the mass term no longer has the factor $\frac{1}{2}$ characteristic of the real field. These types of fields can describe charged vector particles, such as the W^\pm bosons mediating the weak interaction.

The expansion into creation and annihilation operators is analogous to (2.106), up to the fact that due to the gauge invariance for the Maxwell case with vanishing mass, or for the gauge constraint for the massive case of Proca [see the discussion below (2.71)], the number of degrees of freedom is not equal to the number of field components (i.e. four), but three for the massive field and two otherwise. This point leads to subtleties in the quantization that are discussed in Refs. [7–9] to which we direct the reader.

2.2.3.2 Internal symmetries

The Lagrangian (2.103) not only represents two independent scalar fields of the same mass. Indeed, these two fields are the components of a complex number and the Lagrangian is thus invariant under the phase shift

$$\phi \rightarrow \phi' = e^{i\alpha} \phi \quad \text{and} \quad \phi^* \rightarrow \phi'^* = e^{-i\alpha} \phi^*, \quad (2.110)$$

where $\alpha \in \mathbb{R}$ is an arbitrary constant. Expressed in terms of the real and imaginary parts of ϕ , the phase shift (2.110) corresponds to a global internal rotation. It is this invariance that imposes the states 'a' and 'b' to have the same mass.

From Noether's theorem, the existence of such an invariance should translate into the existence of a conserved quantity. To determine it let us work in the limit $\alpha \ll 1$, such that we can expand the exponential in (2.110), leading to the infinitesimal transformation $\phi' = (1 + i\alpha)\phi$. This transformation being independent of the spatial coordinates, x , we have $f_\mu = 0$ in (2.29), and the current

$$\mathcal{J}^\mu := i \frac{\partial \mathcal{L}}{\partial \partial_\mu \phi} \phi - i \frac{\partial \mathcal{L}}{\partial \partial_\mu \phi^*} \phi^*, \quad (2.111)$$

is conserved since in that case, we have $\delta\phi = i\alpha\phi$ and $\delta\phi^* = -i\alpha\phi^*$ (we have removed the irrelevant normalisation factor α in this definition).

The conserved 'charge' is obtained by direct integration, i.e.

$$Q = \int d^3x \mathcal{J}^0 = -i \int d^3x (\pi\phi - \pi^*\phi^*). \quad (2.112)$$

After quantization, this charge leads to the operator

$$\hat{Q} = \int d^3k \left[\hat{\mathcal{N}}_k^{(a)} - \hat{\mathcal{N}}_k^{(b)} \right],$$

(2.113)

which is indeed time independent since its equation of motion, using the form (2.108) of the Hamiltonian, is

$$\frac{d\hat{Q}}{dt} = -i [\hat{Q}, \hat{H}_c] = 0.$$

Equation (2.113) can be easily interpreted by saying that particles associated with the operators \hat{a} and \hat{b} are particles of opposite charges; in reality they are antiparticles of one another.

The complex scalar field is only an example of internal symmetry, in this case of an invariance under transformations of a $U(1)$ group. The associated charge is, in general, called *hypercharge*, and corresponds to the electric charge only in the case where the $U(1)$ group is that of electromagnetism. The different existing particles of which we will make an inventory in Section 2.3, indicate the existence of more symmetries, based on larger symmetry groups.

2.2.3.3 Fermions and anticommutation relations

It is possible to expand the Dirac field (2.63) in a similar way as in the expansion (2.106), with an important difference however: the field operators, and hence those associated to the annihilation and creation of particles, must respect anticommutation relations instead of commutation relations of the type (2.83) or (2.84). The consistency problem related to this point is made explicit in Refs. [7, 8], and is beyond the scope of this presentation. It will be sufficient to know that for fermions we need to replace the Poisson brackets by anticommutation relations

$$\left\{ \hat{\psi}_a(\mathbf{x}, t), \hat{\psi}_b^\dagger(\mathbf{x}', t) \right\} \equiv \hat{\psi}_a(\mathbf{x}, t) \hat{\psi}_b^\dagger(\mathbf{x}', t) + \hat{\psi}_b^\dagger(\mathbf{x}', t) \hat{\psi}_a(\mathbf{x}, t) = \delta_{ab} \delta^{(3)}(\mathbf{x} - \mathbf{x}'), \quad (2.114)$$

and

$$\left\{ \hat{\psi}_a(\mathbf{x}, t), \hat{\psi}_b(\mathbf{x}', t) \right\} = \left\{ \hat{\psi}_a^\dagger(\mathbf{x}, t), \hat{\psi}_b^\dagger(\mathbf{x}', t) \right\} = 0, \quad (2.115)$$

where we used the fact that, from the Lagrangian (2.63), the conjugate momentum of the spinor is given by

$$\pi_\psi \equiv \frac{\delta \mathcal{L}_{\text{Dirac}}}{\delta \dot{\psi}} = i\psi^\dagger. \quad (2.116)$$

Note that these anticommutators, similarly to the commutators for bosonic fields, are taken at equal time. At this point, it is interesting to note that if we place ourselves in the context of field theory at finite temperature (thermal field theory, see e.g., Ref. [10]), then these anticommutation relations are necessary so that the fermions satisfy the Pauli exclusion principle; moreover, they lead to the correct statistical distribution law, i.e. the Fermi–Dirac law (just as the commutation relations for the bosons lead to the Bose–Einstein distribution).

2.3 Classification and properties of the elementary particles

The general properties of elementary particles are gathered in Table 2.1 (see Ref. [11]). When one looks at these properties, one finds structures, such as illustrated, for instance, in Fig. 2.1 showing the particular case of baryons of spin $\frac{3}{2}$ and of positive parity (parity is defined later, here it is only an illustration).

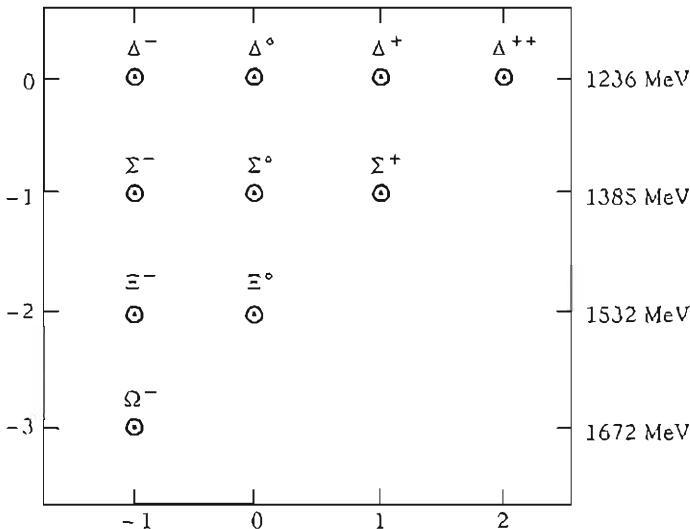


Fig. 2.1 An example of the structure observed between the elementary particles: the baryons of spin $\frac{3}{2}$ and of positive parity.

Table 2.1 Nomenclature of the elementary particles. The families based on quarks (q) have several hundreds of particles identified so far, whereas the leptons and the gauge fields are here presented in an exhaustive way in the context of the standard model. The gauge fields are written in the following way: γ is the photon, the Z^0 and W^\pm bosons are the mediators of the weak interaction, and g^a represent the 8 gluons of the strong interaction.

| Name | Spin | Examples |
|--------------|-------------------|---|
| Hadrons | $n + \frac{1}{2}$ | $p^+, n^0, \Delta, \Lambda, \Sigma, \Omega, \Xi \dots$ |
| | n | $\pi^{0,\pm}, K^{0,\pm}, J/\psi, D^0, B^0, \eta, \dots$ |
| Leptons | $\frac{1}{2}$ | $e^-, \nu_e, \mu^-, \nu_\mu, \tau^-, \nu_\tau$ |
| Gauge fields | 1 | γ, Z^0, W^\pm, g^a |

This kind of diagram has made it possible to understand the existence of a very precise structure, the history of which is discussed in Ref. [12]. In what follows, we will only indicate the result of numerous years of experimental research that led to the so-called ‘standard’ model of particle physics, its theory is the subject of Section 2.6.

2.3.0.4 Exchange bosons

The known particles come in three forms. First, we find the bosons of unit spin, mediating fundamental interactions: they are the gauge bosons, namely the photon γ , which mediates the electromagnetic interaction, the eight gluons g^a , which mediate the

strong nuclear interaction, and the vectors Z^0 and W^\pm . The latter have the property of being massive and are hence described by Proca's Lagrangian (2.109), at least as long as we ignore their interactions. The vectors Z^0 and W^\pm allow the weak nuclear interaction to propagate, W^\pm are in addition electrically charged.

According to the current theoretical ideas, to be complete, one should add to this table the graviton, $h_{\mu\nu}$, massless 'particle' (in the framework of a quantum theory of gravity equivalent to that of the other interactions) of spin 2 responsible for the gravitational interaction (Chapter 1). In the absence of any satisfactory quantum theory of gravity, we do not consider at the moment the graviton as an elementary particle belonging to the standard model.

2.3.0.5 Leptons

Of spin $\frac{1}{2}$, the leptons, among which the electron is the best known, are of negative electrical charge (e^- , μ^- and τ^-) or vanishing (ν_e , ν_μ and ν_τ which are the associated neutrinos). The evolution of these particles is dictated by the Dirac equation, coupled or not to electromagnetism, depending on the case. They are also subject to the weak interaction and can be grouped into three 'doublets'

$$\begin{pmatrix} e^- \\ \nu_e \end{pmatrix}, \quad \begin{pmatrix} \mu^- \\ \nu_\mu \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \tau^- \\ \nu_\tau \end{pmatrix}$$

which have exactly the same properties with respect to this weak interaction. Actually, the only characteristic that allows us to distinguish between these different doublets is the mass of the particles that constitute them. Moreover, among them, only the doublet of the electron is stable, the others, of higher mass, are subject to the weak and electromagnetic interactions, and can decay into, for instance, electrons and electron-neutrinos.

For a long time, the neutrinos have been considered to have a vanishing mass, but some recent experiments and measurements (see Ref. [2] of Chapter 9) have shown that there must be a mixing between the different species of neutrinos, which is only possible for massive particles.

2.3.0.6 Hadrons

Only one category of particles, albeit the most numerous one, is subject to the strong interaction: it is the hadrons. They can be of half-integer spin, such as the proton and the neutron, and then belong to the category of the baryons. The mesons, of integer or vanishing spin, count among them, for instance, the π^0 and its charged partners π^\pm , which are described by the Klein-Gordon equation for the real scalar fields (as the π^0) or the complex (for the π^\pm).

The surprising profusion of the hadrons, as well as the observed properties of symmetry between the different elements of the families sharing some properties (spin and parity, for instance, as in the illustration of Fig. 2.1) have led to the understanding that these particles, unlike what we currently think of gauge bosons and leptons, are not elementary particles, but are constituted of 'quarks'.³ The latter are for the

³According to the legend, this word was proposed by Murray Gell-Mann (Nobel Prize 1969) who would have read it in the book 'Finnegan's Wake' by James Joyce.

moment seen as effectively elementary, of spin $\frac{1}{2}$, and of electric charge $+\frac{2}{3}$ for the quarks u , c and t , and $-\frac{1}{3}$ for the quarks d , s and b . We can also group them into doublets for the weak interaction, with a similar structure to the one for the leptons, more precisely

$$\begin{pmatrix} u \\ d \end{pmatrix}, \quad \begin{pmatrix} c \\ s \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} t \\ b \end{pmatrix}.$$

This description, which is satisfying from the theoretical point of view, currently does not have any generally accepted explanation (even if many models exist).

The names of the quarks themselves come from this structure: at the time where only the first family was experimentally accessible, the chosen names were simply *up* (u) and *down* (d). Then, a third quark was needed to be introduced, which was considered as *strange* (s). The second family was then completed by the *charm* (c). In the same philosophy, when the presence of the fifth quark started to appear, the name of *beauty* (b) was proposed, soon replaced by *bottom*, allowing the first family to be reproduced with the *top* (t).

Just as for the leptons, the only difference between the quark families comes from their masses, as well as the fact that only the members of the first family are stable. The ones from the second and third families decay much faster than their leptonic counterparts since they do so through the strong interaction instead of the weak interaction, the latter having a much weaker amplitude, as its name suggests.

All these particles have been produced in accelerators, but have never been observed alone, which can be explained by the notion of confinement of strong interactions. This is how it is only possible to observe states containing three quarks, baryonic states of half-integer spin, and states composed of one quark and one antiquark, of integer spin: a meson. This point is not yet completely understood theoretically.

Table 2.1 summarizes the nomenclature used to describe the diversity of elementary particles. The origin of the mass of all these particles, together with the important disparities observed between them, still pose many questions that are not all yet resolved.

As we can see, the known particles reflect, or seem to reflect, some elements of symmetry. These are described by group theory, to which we now turn.

2.4 Internal symmetries

Contrary to leptons, which all appear in doublets, hadrons are found in numerous forms, suggesting that they are not elementary. In reality, using a much smaller number of elementary particles, in this case quarks, and placing them in a relevant way in group representations, it is actually possible to combine them so that the other representations are perceived as new elementary particles. The role of group theory is precisely to lead us to this conclusion.

2.4.1 Group theory

2.4.1.1 Definitions

A group \mathcal{G} is defined as a collection of objects $\{g_i\}$, where the index i is not necessarily discrete, satisfying the following rules:

- There exists a composition law, denoted by \circ , which can combine any two elements of the group, the resulting combination being an element of the group: if $g_1 \in \mathcal{G}$ and $g_2 \in \mathcal{G}$, then $g_3 = g_1 \circ g_2 \in \mathcal{G}$. Furthermore, this law is associative, i.e. $g_1 \circ (g_2 \circ g_3) = (g_1 \circ g_2) \circ g_3 = g_1 \circ g_2 \circ g_3$.
- There exists a neutral element, denoted by 1 , such that $\forall g \in \mathcal{G}, 1 \circ g = g \circ 1 = g$.
- Each element $g \in \mathcal{G}$ is associated with an inverse element, denoted by g^{-1} , satisfying $g \circ g^{-1} = g^{-1} \circ g = 1$.

Note that the law \circ is not necessarily commutative: when it is, i.e. if $\forall g_1, g_2 \in \mathcal{G}, g_1 \circ g_2 = g_2 \circ g_1$, the group is then said to be *Abelian*. In the opposite case, the group is usually said to be *non-Abelian*.

The simplest group one can think of, besides the trivial group that only contains one element, is the discrete group Z_2 that contains two elements, denoted by 1 (which is the identity) and -1 with the ordinary multiplication law. This is an Abelian group for which each element is its own inverse. For discrete groups, one can construct the complete table of all possible operations.

2.4.1.2 Lie Groups

Another well-known example of a group is that of translations that transforms a point in space x into $x + a$. In this case, any element of the group can be directly associated with a translation vector a : the composition law is vector addition, the inverse element of a is simply $-a$, and the identity is the translation by the null vector 0 . It is again an Abelian group.

Such a group contains an infinite number of elements, which can all be identified by the knowledge of three real numbers (the three coordinates of the translation vector). More generally, if it is possible to describe the elements of a group by a set of numbers varying in a continuous interval (but not necessarily infinite), it is then a Lie group. A Lie group is also a manifold.

The best-known non-Abelian group is probably that of rotations in three dimensions. Each rotation is determined by, for instance, the three Euler angles, or equivalently, a unit vector (rotation axis) and the angle of rotation around this rotation axis. The composition of two rotations is equivalent to a third rotation, which makes it a group (since a rotation can always be inverted by performing the same rotation in the other way, and so a rotation of angle zero is the identity), but the order in which the rotations are performed is crucial and this group is not Abelian. This group, denoted by $SO(3)$, is identical to the group formed by the orthogonal 3×3 matrices with unit determinant. We can generalize it to the group $SO(n)$, which contains the orthogonal $n \times n$ matrices with unit determinant.

Each group is associated with an invariance. For instance, in the case of a rotation, the norm of a vector on which the rotation has been performed remains unchanged during the transformation. A rotation in a generalized Minkowski space, i.e. having p spatial dimensions and q time-like dimensions, must keep the form $u_\mu u^\mu$ invariant; it contains not only rotations in the p -dimensional space, but also pure Lorentz transformations (some ‘boosts’). This group, generalizing $SO(n)$ for arbitrary metric signatures, describes p space-like coordinates and q time-like ones, and is denoted by $SO(p, q)$. The so-called Lorentz group, is then simply $SO(3, 1)$. The translation group

can be associated with the Lorentz group to form the Poincare group on which special relativity is based (see Chapter 1).

2.4.2 Generators

2.4.2.1 Infinitesimal transformations

An infinitesimal transformation, T_ϵ , of a Lie group is a transformation close to the identity and hence can always be written as $T_\epsilon \simeq 1 + i\epsilon_a \lambda^a$. This defines the generators λ^a , which are Hermitian matrices and depend on the nature of the performed transformation. We suppose here that the Lie group is described by p parameters, i.e. $a = 1, \dots, p$.

A finite transformation can be obtained from an infinitesimal transformation by exponentiation. Such a finite transformation can always be written as the composition of a large number, tending to infinity, of infinitesimal transformations. If α represent the p necessary numbers for the finite transformation, then we can always write $\epsilon = \alpha/N$, where $N \gg 1$, and performing N transformations will indeed lead to the finite transformation. In other words, we will have

$$T_\alpha = \lim_{N \rightarrow \infty} \left(1 + i \frac{\alpha}{N} \cdot \lambda \right)^N = e^{i\alpha \cdot \lambda}, \quad (2.117)$$

which is the general form of a finite transformation of the group.

2.4.2.2 Lie Algebra

In order for $T_\alpha \circ T_\beta$ to be a group transformation, T_γ , for instance, one should have the property

$$e^{i\alpha \cdot \lambda} e^{i\beta \cdot \lambda} = e^{i\gamma \cdot \lambda}. \quad (2.118)$$

For infinitesimal transformations this gives to second order in the expansion of the exponentials

$$e^{i\alpha \cdot \lambda} e^{i\beta \cdot \lambda} = 1 + i(\alpha_a + \beta_a) \lambda^a - \frac{1}{2} (\alpha_a \alpha_b + \beta_a \beta_b + 2\alpha_a \beta_b) \lambda^a \lambda^b,$$

the logarithm of which, also expanded to second order [using the relation $\ln(1 + \epsilon) = \epsilon - \frac{1}{2}\epsilon^2 + \dots$], should be proportional to λ^a . We find explicitly

$$i(\alpha_a + \beta_a) \lambda^a + \frac{1}{2} \alpha_a \beta_b [\lambda^a, \lambda^b] = i\gamma_c \lambda^c.$$

This equality is in general only satisfied if the commutator of the λ satisfies

$[\lambda^a, \lambda^b] = i\Gamma^{ab}_c \lambda^c.$

(2.119)

This relation defines the algebra associated with the Lie group, i.e. the set of operators of the form $\alpha_a \lambda^a$. The numbers Γ^{ab}_c are the structure constants of the group, and the λ^a form its algebra. Note that since the generators are Hermitian operators, the structure constants are real numbers.

2.4.2.3 Representations

In order to be able to classify the particles in terms of symmetry groups, we need to sort them depending on the representations of the group of the considered invariance. An n -dimensional representation being a set $\{\mathcal{R}^a\}$ of $n \times n$ matrices satisfying the commutation relations (2.119) of the algebra, the fields that describe the particles of a given representation will be gathered in multiplets $\Phi = \{\Phi_a\}$, $a = 1, \dots, n$ on which the action of the group is

$$\Phi \longrightarrow (T_\alpha \Phi)_a \equiv \left(e^{i\alpha_c \mathcal{R}^c} \right)_a^b \Phi_b. \quad (2.120)$$

It is usual to say that Φ_a transforms under the representation \mathcal{R}^a , or more directly, that Φ_a is ‘in the representation \mathcal{R}^a ’. If two representations are related to each other by a similarity transformation, i.e.

$$\tilde{\mathcal{R}}^a = U \mathcal{R}^a U^\dagger,$$

where U is unitary ($U^\dagger = U^{-1}$), then we say that they are equivalent. In other words, for all practical purposes, it is the same representation, such that the similarity transformation does not provide any complementary information. If there is an invariant subspace in the representation, i.e. if the action of any element of the group on the vectors of this subspace leads to an element of the same subspace, then the representation is said to be reducible. In the opposite case, it is called *irreducible*. Only the irreducible representations are used to classify the particles in the representations of the invariance groups.

Moreover, if $\{\mathcal{R}^a\}$ forms a representation, then so does $\{-(\mathcal{R}^a)^*\}$, which is called the ‘conjugate representation’, [one can easily be convinced that these matrices also respect the algebra closure relation (2.119) since the structure constants of the group are real]. If a representation and its conjugate are equivalent then we are dealing with a so-called real representation. Otherwise, we say the representation is complex. In order for a group to be able to potentially reproduce the internal symmetries of particle physics, it is often required that it has some complex representations in which fermions can be placed in a simple way.

The representations that are the most relevant for quantum physics are those that preserve the norm defined by

$$|\Phi|^2 \equiv \Phi^\dagger \Phi,$$

of the multiplet Φ . In order for the norm to be conserved during an arbitrary transformation \mathcal{R}^a , we see that we need to have

$$(\Phi)^\dagger \Phi = \Phi'^\dagger \Phi' = (\Phi)^\dagger (T_\alpha)^\dagger T_\alpha \Phi,$$

for all transformations T_α . We should therefore require that $(T_\alpha)^\dagger T_\alpha = 1$, i.e. that it is a unitary transformation. A representation satisfying this condition is a unitary representation.

Finally, the adjoint representation of a Lie group is formed using its structure constants. It is a representation of the same dimension as the order of the group, in which the matrix elements of \mathcal{R}^a are

$$(\mathcal{R}^b)_c^a = \Gamma^{ab}_c. \quad (2.121)$$

2.4.2.4 Cartan classification

Besides the space-time symmetries (Poincaré group), the invariance groups that are useful from the point of view of the classification of particles are the compact ones, i.e. for which the parameters vary on a finite interval, as if we had to deal with generalized rotations in an internal space (and in many cases, it is actually precisely the required invariance). Moreover, we wish to classify the particles along 'fundamental' invariances, i.e. with no substructure (we will see later that only one coupling constant comes into play in these groups, which makes them even more interesting). This is why we are interested in simple groups: if \mathcal{H} is a subgroup of \mathcal{G} and if $\forall g \in \mathcal{G}$ and $\forall h \in \mathcal{H}$, the element $g \circ h \circ g^{-1}$ is still in \mathcal{H} , then \mathcal{H} is an invariant subgroup of \mathcal{G} . \mathcal{G} is simple if there is no invariant subgroup in \mathcal{G} , and semi-simple if it does not contain any Abelian invariant subgroup. One can show that a compact and semi-simple group is thus either forced to be simple or is the direct product of two or more simple groups. Using this philosophy, studying simple Lie groups is therefore sufficient to classify all semi-simple compact groups. Note also that the groups with structure $\mathcal{S} \times [\mathrm{U}(1)]^n$ with \mathcal{S} a semi-simple group and n an arbitrary integer, which are called reducible groups, have representations that can all be decomposed as a product of irreducible representations; this property makes them especially interesting for the classification of elementary particles.

Since the simple Lie groups are also subject to the constraint (2.119), they can be classified in terms of their structure constants. This classification was performed by E. Cartan at the beginning of the twentieth century. He proved that the simple Lie groups could be classified into 4 large categories, namely $\mathrm{SO}(2n)$, $\mathrm{SO}(2n+1)$, $\mathrm{SU}(n+1)$, and $\mathrm{Sp}(2n)$, to which one should add the so-called exceptional groups, which can not be classified into any of these categories, called G_2 , F_4 , E_6 , E_7 and E_8 . For the four infinite sequences, n is the rank of the group, i.e. the maximal number of generators that can be simultaneously diagonalised. For the exceptional groups, the rank is given by the index in the notation. For each group, the order, i.e. the number of generators, is indicated in Table 2.2.

Table 2.2 Order (number of generators) or dimension of the simple Lie groups of rank n according to the Cartan classification (the index of the exceptional groups is the rank).

| Group | | $N_{\text{generators}}$ |
|---------------------|--------------|-------------------------|
| $\mathrm{SU}(n+1)$ | $(n \geq 1)$ | $n(n+2)$ |
| $\mathrm{SO}(2n+1)$ | $(n \geq 2)$ | $n(2n+1)$ |
| $\mathrm{Sp}(2n)$ | $(n \geq 3)$ | $n(2n+1)$ |
| $\mathrm{SO}(2n)$ | $(n \geq 4)$ | $n(2n-1)$ |
| G_2 | | 14 |
| F_4 | | 52 |
| E_6 | | 78 |
| E_7 | | 133 |
| E_8 | | 248 |

The groups of the four infinite sequences can be understood using their fundamental representation, i.e. the one of smallest dimension (from which, by direct product, all the other representations can be constructed). The set of all complex unitary matrices of dimension $n \times n$ forms the group $U(n)$ ('U' for 'unitary'), the subgroup restricted to the unit determinant matrices is $SU(n)$ ('S' for 'special'). It is the set of transformations that leave invariant the inner product of two vectors of a complex space of dimension n . The transformations of this group leave invariant, among other things, the norm of the complex vectors and can be used in quantum mechanics to assure the conservation of probabilities.

The sequence $SO(n)$ is constructed from the matrices that leave invariant the inner product of two real vectors. They are orthogonal matrices (hence the 'O') with unit determinant (special) of dimension $n \times n$. Finally, the group $Sp(2n)$ regroups all the matrices M of dimension $2n \times 2n$ that leave invariant the antisymmetric matrix

$$S = \begin{pmatrix} 0 & 1 & & & \\ -1 & 0 & & & \\ & & \ddots & & \\ & & & 0 & 1 & \\ & & & -1 & 0 & \\ & & & & & \ddots & \\ & & & & & & 0 & 1 \\ & & & & & & -1 & 0 \end{pmatrix}.$$

These matrices therefore satisfy $M^T S M = S$, where M^T is the transpose matrix of M . Thus, the antisymmetric and quadratic form $y^T S x$ for two real vectors of dimension $2n$ is conserved. The matrices M defined in this way are called the symplectic matrices, thus the name of the group.

The only transformations that allow us to conserve probabilities in quantum mechanics are those that are unitary. Can we hence deduce that only the groups of the type $SU(n)$ can play a role in physics, and that hence all theories should be based on such a group? Unfortunately, the answer to this question is negative: some representations of the groups classified by Cartan are unitary, as well as respecting the rules of their classification. We can therefore exclude a group as a possible internal invariance only if we can show that it has no unitary representation. This is, for instance, the case of $SO(11)$, which is hence excluded as a grand unification group candidate. This is why it is indispensable to study in detail the representations of the possible Lie groups.

2.5 Symmetry breaking

Not all the symmetries existing in Nature are manifest, and some underlying ones are not immediately visible at low energy. This is, for instance, the case for the doublets of the type (e^-, ν_e) . If these particles really belong to the same representation of the gauge group of the weak interaction, why do they not have the same characteristics as well (mass, electric charge, for instance)? The answer to this question relies on the mechanism that enables us to break symmetries, and in particular the gauged ones.

2.5.1 Gauge group

We are generally confronted with ‘gauge’ invariances, i.e. local invariances. This leads to the existence of a gauge boson, through which some interactions are performed: *the presence of a gauged (local) symmetry in particle physics, as in gravity, implies the existence of an interaction.*

2.5.1.1 Local symmetries

The simplest illustration of a local symmetry is given by electromagnetism. Its action is invariant under the transformation of the group U(1), which leads to the prediction of the existence of the photon, i.e. the term (1.94) of Chapter 1 in the total action.

Consider a complex scalar field, $\phi(x^\mu) \in \mathbb{C}$, invariant under the phase transformations $\phi \rightarrow \phi' = e^{ic_\phi \alpha} \phi$, where c_ϕ is a real constant. If only one field is present, c_ϕ can be put equal to unity, but if there are several fields, the ratio of the factors c gives the ratio of the ‘charges’ of the different fields under the U(1) transformation. If the Lagrangian of this field is of the form (2.103), it remains unchanged under the transformation as long as α is a constant. However, if $\alpha = \alpha(x^\mu)$, the kinetic term is not invariant and transforms as

$$\begin{aligned} (\partial_\mu \phi)^* \partial^\mu \phi &\rightarrow (\partial_\mu \phi')^* \partial^\mu \phi' = [(\partial_\mu - ic_\phi \partial_\mu \alpha) \phi^*] [(\partial^\mu + ic_\phi \partial^\mu \alpha) \phi] \\ &= (\partial_\mu \phi)^* \partial^\mu \phi + ic_\phi \mathcal{J}^\mu \partial_\mu \alpha + c_\phi^2 |\phi|^2 \partial_\mu \alpha \partial^\mu \alpha, \end{aligned} \quad (2.122)$$

where the conserved current $\mathcal{J}^\mu = -i(\phi^* \partial^\mu \phi - \phi \partial^\mu \phi^*)$ is given by (2.111). We can note here that this term, proportional to the conserved current, leads, after integration by parts, to a surface term in the action and thus has no physical effects. Despite this fact, the action is still not invariant since the last term is remaining.

2.5.1.2 Gauge boson

The form (2.122) suggests the replacement of the partial derivative by a gauge covariant derivative by introducing a new vector field C^μ . If we define

$$D^\mu \phi \equiv (\partial^\mu - igc_\phi C^\mu) \phi, \quad (2.123)$$

with g an arbitrary coupling constant, we find that this term transforms as

$$D^\mu \phi \rightarrow (D^\mu \phi)' = \left[\partial^\mu + igc_\phi \left(C'^\mu + \frac{1}{g} \partial^\mu \alpha \right) \right] \phi.$$

We can therefore simply impose C^μ to transform as

$$C^\mu \rightarrow C'^\mu = C^\mu - \frac{1}{g} \partial^\mu \alpha, \quad (2.124)$$

in order for the action to be again invariant, whatever the value of c_ϕ . Since the action (1.94) with the Minkowski metric $\eta^{\alpha\beta}$ is also invariant under the transformation (2.124), we can therefore consider the more general following action

$$S_s = \int d^4x \left[-(D_\mu \phi)^* D^\mu \phi - V(|\phi|) - \frac{1}{4} G_{\mu\nu} G^{\mu\nu} \right], \quad (2.125)$$

where $G_{\mu\nu} = \partial_\mu C_\nu - \partial_\nu C_\mu$, and we have generalized the potential of the scalar field for it to be an a priori arbitrary function of the field norm $|\phi| \equiv \sqrt{|\phi|^2}$. The action (2.125) is the basis of the *scalar electrodynamics*, an example of gauge theory.

We should note that (2.125) does not have any mass term for the vector field C^μ . This comes from the fact that such a term, described by the Lagrangian of Proca (2.71), is not invariant under the transformation (2.124).

The action (2.125) may be generalized to the case where the gauge group is not Abelian, this is what is done in the context of the standard model of Section 2.6. Note already here that the gauge bosons introduced in the framework of non-Abelian theories can neither be massive, for the same reason that a mass term à la Proca is not gauge invariant.

2.5.1.3 Fermions

In a similar way, the action for the fermions (2.67) is made invariant under the local transformation $\psi \rightarrow \psi' = e^{ic\alpha}\psi$ by replacing the partial derivative by the gauge covariant derivative

$$D_\mu \psi = (\partial_\mu + igc_i C_\mu)\psi. \quad (2.126)$$

We can also see that an extra coupling term with the scalar field can be introduced. Finally, the action for the fermions, can be written in a general form as

$$S_f = - \int d^4x \left[\frac{1}{2} \bar{\psi} i \gamma_\mu \overleftrightarrow{D}^\mu \psi + \bar{\psi} (m_f + f|\phi|) \psi \right], \quad (2.127)$$

with f an arbitrary constant, which, in addition of (2.125), allows us to describe a theory containing a fermion and a complex scalar field coupled via an intermediate vector. We can also see that the fermion can have a purely massive term in addition to its coupling with the scalar field, while respecting the invariance.

2.5.2 Higgs mechanism

The Higgs mechanism allows us to build a theory invariant under the transformations of a given group, while this invariance is not manifest. In other words, it allows us to impose an invariance with some properties to theories that do not seem to have it, as is, for instance, the case for theories having massive gauge bosons.

2.5.2.1 Higgs potential

We consider a complex scalar field ϕ and a gauge field A^μ , the dynamics of which is described by the action (2.125). For technical reasons that we will not discuss here, (of renormalization, see, for instance, Ref. [7]), the potential $V(|\phi|)$ cannot introduce powers of $|\phi|$ higher than 4. We generally choose a quadratic term, i.e. we postulate a mass for the scalar field, and a quartic term, in $\lambda|\phi|^4$. In the case of the Higgs mechanism, the chosen potential is given by

$$V_{\text{Higgs}}(\phi) = \lambda (|\phi|^2 - \eta^2)^2, \quad (2.128)$$

depicted in Fig. 2.2. The parameter η , having the dimension of a mass, is called a ‘vacuum expectation value’ (VEV).

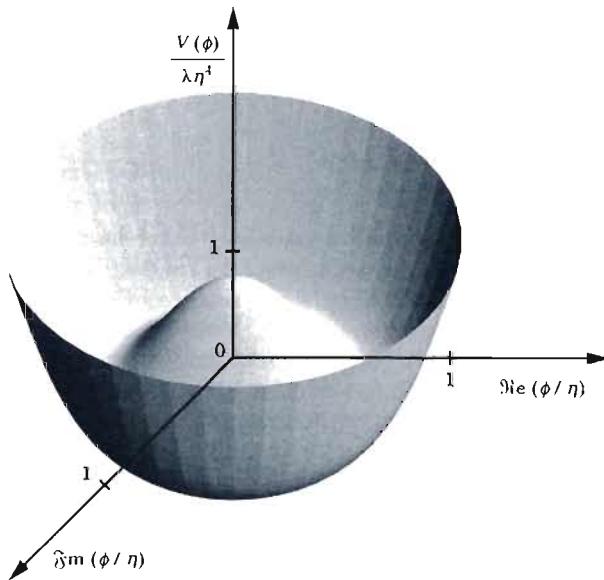


Fig. 2.2 The Higgs potential as a function of the components (real and imaginary parts) of the Higgs field. The circle $|\phi|^2 = \eta^2$ indicates the set of possible minima.

For the theory to make sense, one should first define the vacuum. It can be obtained, at the classical level, by demanding the energy, i.e. the Hamiltonian, to be a minimum. A simple generalization of (2.88) gives in that case

$$\mathcal{H}(x, t) = \frac{1}{2} [\dot{\phi}^2 + (\nabla\phi)^2 + 2V(\phi)], \quad (2.129)$$

which is the sum of two positive-definite terms and of the potential. As a consequence, the minimal energy is reached when the field is invariant with respect to the transformation of space and time, so that $\dot{\phi} = \nabla\phi = 0$. Furthermore, the potential should also be minimized, which leads to

$$\frac{dV}{d\phi} = 0 \implies \phi^* (|\phi|^2 - \eta^2) = 0. \quad (2.130)$$

As illustrated in Fig. 2.2, one of the solutions, here $\phi = 0$, represents a local maximum of the potential, thus it is an unstable state of the theory. The second solution, $|\phi| = \eta$, on the other hand, is the absolute minimum of the potential. This is the value the field will hence naturally take and that describes the vacuum state of the theory (thus the name of the parameter η).

2.5.2.2 Goldstone boson

The expansion of the field ϕ around its VEV is

$$\phi = \left[\eta + \frac{1}{\sqrt{2}} h(x, t) \right] e^{i\theta(x, t)/\sqrt{2}\eta}, \quad (2.131)$$

where h and θ are two real fields. Noting $m_h \equiv 2\sqrt{\lambda}\eta$, the potential for h and θ can be written as,

$$V(h, \theta) = \frac{1}{2}m_h^2 h^2 + \sqrt{2}\lambda\eta h^3 + \frac{\lambda}{4}h^4, \quad (2.132)$$

and is independent of θ . This can easily be understood by examining the potential: the fluctuations h and θ around the expectation value η of the field correspond to fluctuations either along the radial direction (h), thus changing the energy of the configuration, or along the circle of the minima (θ), which by definition do not cost any energy. The field h is called the Higgs field, whereas θ is the Goldstone boson.

2.5.2.3 Masses of the bosons and fermions

The degree of freedom corresponding to the Goldstone boson is not physical. Actually it is always possible to choose a gauge in which it is suppressed, since the only place where it comes in is in the kinematic term of the field ϕ , which can be expanded as

$$(D_\mu\phi)^* D^\mu\phi = \frac{1}{2}(\partial h)^2 + \left(\eta + \frac{1}{\sqrt{2}}h\right)^2 \left(gc_\phi C + \frac{1}{\sqrt{2}\eta}\partial\theta\right)^2. \quad (2.133)$$

Choosing the gauge in which $C_\mu \rightarrow C_\mu + (\partial_\mu\theta)/(\sqrt{2}\eta gc_\phi)$, the field θ completely disappears.

After expanding the kinematic part (2.133) of the scalar field ϕ , one can see appearing a term similar to Proca's one (2.70) for the vector field C_μ . This implies that the field has gained a mass during this process. After identification, we find $M_C = \sqrt{2}gc_\phi\eta$.

The appearance of a massive vector field can be explained by the loss of the Goldstone boson. Its corresponding degree of freedom cannot have completely disappeared, and it has actually been completely absorbed ('eaten', as it is sometimes written) by the gauge field. A massless vector field has indeed only two degrees of freedom, corresponding in the classical theory to the two transverse polarization modes, either circular or linear. On the other hand, a massive vector field evolves such that its angular momentum has a vanishing projection, and therefore has three possible modes.

The coupling of the scalar ϕ to the fermion in (2.127) leads to a modification of the fermion's mass, which is increased (or decreased, depending on the sign of the coupling constant f), by $\Delta m_f = f\eta$. Thus, if the fermion was by symmetry initially massless, as is the case in the electroweak and strong standard model, it will thus acquire a mass, in the same way as the gauge field. This is why fermions of extremely different masses can be placed in the same multiplet. We will see later, in the context of the standard model of the electroweak and strong interactions, how this process is applied in a precise case, and is checked experimentally.

2.5.3 Non-Abelian case

The Higgs mechanism can easily be generalized to an arbitrary group \mathcal{G} , which could be non-Abelian, but that we suppose is simple. We now consider a field in a representation

of dimension n , i.e. a column vector ϕ composed of n real scalar fields ϕ_a , with $a = 1, \dots, n$. The VEV is now a vector

$$\eta = \langle 0 | \phi | 0 \rangle = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}, \quad (2.134)$$

with $|0\rangle$ the vacuum state of the scalar fields. The individual VEVs, $\eta_a = \langle 0 | \phi_a | 0 \rangle$, satisfy $\sum_a \eta_a^2 = \eta^2$. The potential of ϕ is always of the form (2.128), with $|\phi| = \sum_a \phi_a^2$, so that once again we have

$$\frac{\partial V(\phi)}{\partial \phi_a} \Big|_{\phi=\eta} = 0,$$

and we define the quantum fields $\hat{\phi}_a \equiv \phi_a - \eta_a$. The mass term then takes the general form

$$\mathcal{L}_{\text{mass}} = -\frac{1}{2} \frac{\partial^2 V}{\partial \phi_a \partial \phi_b} \hat{\phi}_a \hat{\phi}_b.$$

If we perform a global infinitesimal transformation with parameter $\alpha_c \ll 1$, the potential transforms as $V(\phi) \rightarrow V(\phi) + \delta V(\phi)$, with

$$\delta V(\phi) = \frac{\partial V}{\partial \phi_a} \delta \phi_a = i \frac{\partial V}{\partial \phi_a} (\alpha_c \mathcal{R}^c)^b{}_a \phi_b,$$

where we assume that ϕ transforms under the given representation with the matrices \mathcal{R}^c .

For the potential to be invariant under the previous transformation, we need $\delta V(\phi) = 0$, implying that

$$\frac{\partial V}{\partial \phi_a} (\mathcal{R}^c)^b{}_a \phi_b = 0.$$

Differentiating with respect to ϕ_d for $\phi = \eta$ (for which the first partial derivative of the potential cancels), leads to

$$\frac{\partial^2 V}{\partial \phi_a \partial \phi_d} \Big|_{\phi=\eta} (\mathcal{R}^c \eta)_a = 0.$$

There are now two possible ways to make these terms vanish, according to whether or not $(\mathcal{R}^c \eta)_a$ vanishes. The set of the generators satisfying $(\mathcal{R}^c \eta)_a = 0$ forms a subgroup \mathcal{H} of \mathcal{G} under which the vacuum is still invariant. Thus, the scheme of the symmetry breaking is $\mathcal{G} \rightarrow \mathcal{H}$.

The expansion (2.131) around the VEV can be generalized to

$$\phi = \exp \left[i \sum_{i=p+1}^n \theta_i(x) \mathcal{R}^i \right] \begin{pmatrix} \eta_1 + h_1(x) \\ \vdots \\ \eta_p + h_p(x) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad (2.135)$$

assuming that only the p first values of ϕ are non-vanishing. Just as in the Abelian case, the $n - (p + 1)$ Goldstone bosons $\theta_i(x)$ which come in the expansion (2.135) are not physical if we have to do with a local invariance, and their corresponding degrees of freedom produce mass terms for the associated gauge bosons.

We can now change gauge using precisely the Goldstone fields, in a similar way as was done in the Abelian case. The choice (2.131) here becomes

$$\phi \longrightarrow \phi' = \exp \left[-i \sum_{i=p+1}^n \theta_i(x) \mathcal{R}^i \right] \phi(x),$$

so that ϕ becomes

$$\phi' = \begin{pmatrix} \eta_1 + h_1(x) \\ \vdots \\ \eta_p + h_p(x) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad (2.136)$$

which defines the p physical Higgs fields of the theory.

Under a general transformation with parameters α_a , the gauge bosons $C_{a\mu}$ transform as the generalization of (2.124) to the non-Abelian case, more precisely

$$C_{a\mu} \longrightarrow C'_{a\mu} = C_{a\mu} - \frac{1}{g} \partial_\mu \alpha_a + \Gamma^{bc}{}_a \alpha_b C_{c\mu}, \quad (2.137)$$

where g gives the coupling of the group and the $\Gamma^{bc}{}_a$ its structure constants. We should stress that there is only one coupling constant, valid for all the fields submitted to the required invariance. This property is only valid for a simple group or a simple group risen to any power with a transverse symmetry, which allows us to identify the different groups.

In the case of the Higgs mechanism, we use this relation for the transformations that only bring into play the $n - (p + 1)$ Goldstone fields θ_i , $i = p + 1, \dots, n$ instead of the general parameters α_a . To first order in θ_i , the relation (2.137) gives

$$C'_{i\mu} = C_{i\mu} - \frac{1}{g} \partial_\mu \theta_i + \mathcal{O}(\theta^2).$$

Substituting this expression, together with (2.136) in the kinetic term of ϕ , which can be written as $\frac{1}{2} D_\mu \phi \cdot D^\mu \phi$, we find the mass terms of the gauge bosons. In Section 2.6.2, we will see an illustration of this symmetry breaking in the context of the standard model, which we will now study.

2.6 The standard model $SU(3)_c \times SU(2)_L \times U(1)_Y$

The standard model of non-gravitational interactions accounts for all the interactions mentioned in Section 2.3 classified depending on the representations of the groups

$SU(3)_c$ for the strong interactions, and $SU(2)_L \times U(1)_Y$ for the electroweak ones. This last symmetry is broken by the Higgs mechanism described in Section 2.5.2 along the following symmetry-breaking process

$$SU(2)_L \times U(1)_Y \rightarrow U(1)_{\text{elec}},$$

leaving the photon massless, as observed.

2.6.1 The strong interaction (QCD)

The description of the strong interaction is based on an invariance under the transformations of the group $SU(3)$. Since this group has 8 generators, its joint representation, in which the gauge bosons G_μ^a are placed, is of dimension 8: $a = 1, \dots, 8$. The leptons are not subject to these interactions, and thus appear as singlets of $SU(3)$, which do not couple to the gauge boson. The quarks exist in three possible states, and are thus placed in a three-dimensional representation, denoted by 3. The quarks are therefore represented in column vectors with these three elements,

$$Q \equiv \begin{pmatrix} u_r \\ u_g \\ u_b \end{pmatrix}, \quad \begin{pmatrix} d_r \\ d_g \\ d_b \end{pmatrix}, \quad \begin{pmatrix} s_r \\ s_g \\ s_b \end{pmatrix}, \quad \begin{pmatrix} c_r \\ c_g \\ c_b \end{pmatrix}, \quad \begin{pmatrix} t_r \\ t_g \\ t_b \end{pmatrix}, \quad \begin{pmatrix} b_r \\ b_g \\ b_b \end{pmatrix}. \quad (2.138)$$

The index is called colour,⁴ which has nothing to do with any actual coloured appearance. Instead, this terminology comes from an analogy: due to the fundamental properties of the strong interaction, a quark can never be observed alone but only in pairs of quark-antiquarks, or in sets of three quarks. In the latter combinations, the total colour cancels, in the same way as it is possible to combine the three fundamental colours to give white, which can be seen as the absence of any colour. This is why this $SU(3)_c$ group is called the colour group.

The covariant derivative of each quark Q , with respect to the strong interaction is given by

$$D_\mu^{(3)} Q = (\partial_\mu - ig_3 G_{a\mu} \lambda^a) Q, \quad (2.139)$$

where the λ^a are 8 matrices forming the algebra of $SU(3)$. We often take the Gell-Mann matrices (see, for instance, Refs. [7, 13] for more details) which satisfy the appropriate commutation relations of $SU(3)$, i.e. (2.119) with the structure constants of $SU(3)$, denoted as f_{abc}^{ab} . The coupling constant g_3 of the group measures the intensity of the interaction. Experimentally, it has been measured that this coupling constant has a higher numerical value than those for the other interactions, thus the name of strong interaction. By analogy with the colour states, this interaction is also called quantum chromodynamics ‘QCD’.

The Lagrangian of QCD takes the form

$$\mathcal{L}_{\text{QCD}} = - \sum_{\text{quarks}} \bar{Q} \gamma_\mu D_{(3)}^\mu Q - \frac{1}{4} G_{a\mu\nu} G^{a\mu\nu}, \quad (2.140)$$

where $a = 1, \dots, 8$ and $G_{a\mu\nu} \equiv \partial_\mu G_{a\nu} - \partial_\nu G_{a\mu} + if_{abc}^{bc} G_{b\mu} G_{c\nu}$.

⁴The index that takes the value r, g and b represents the ‘fundamental’ colours red, green and blue.

2.6.2 Electroweak interaction

The electroweak interaction combines the properties of the weak interaction, coming from the invariance under transformations of the group $SU(2)_L$, and of electromagnetism, based on $U(1)_{\text{elec}}$, brought together as a unique interaction. All particles are combined in singlets or doublets of $SU(2)_L$.

2.6.2.1 Quarks

The quarks of each generation are supposed to be equivalent, so that the doublets are

$$\begin{pmatrix} u_L \\ d_L \end{pmatrix}, \quad \begin{pmatrix} c_L \\ s_L \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} t_L \\ b_L \end{pmatrix}, \quad (2.141)$$

where each fermion appears as an eigenstate of chirality with the definition

$$\psi_L \equiv \left(\frac{I - \gamma^5}{2} \right) \psi, \quad \psi_R \equiv \left(\frac{I + \gamma^5}{2} \right) \psi, \quad (2.142)$$

where the matrix γ^5 is defined as

$$\gamma^5 \equiv i\gamma^0\gamma^1\gamma^2\gamma^3 = -\frac{i}{4}\epsilon^{\mu\nu\alpha\beta}\gamma_\mu\gamma_\nu\gamma_\alpha\gamma_\beta = \begin{pmatrix} -I & 0 \\ 0 & I \end{pmatrix}.$$

One can explicitly check that $(\gamma^5)^2 = I$, so that the operators $\mathcal{P}_{R,L} \equiv \frac{1}{2}(I \pm \gamma^5)$ are projectors orthogonal to each other, i.e. they satisfy $\mathcal{P}_{R,L}^2 = \mathcal{P}_{R,L}$ and $\mathcal{P}_R \mathcal{P}_L = 0$. Another useful property is to do with the anticommutation relations: γ^5 anticommutes with all γ^μ , or

$$\{\gamma^5, \gamma^\mu\} = 0, \quad (2.143)$$

relation that will be very useful in the following sections.

Finally, the quarks with right chirality are defined as singlets of $SU(2)_L$:

$$u_R, d_R, c_R, s_R, \quad \text{and} \quad t_R, b_R.$$

2.6.2.2 Chirality

The projectors $\mathcal{P}_{R,L}$ on the state of definite chirality⁵ come into play as soon as we are interested in the properties of the spinor with respect to the Lorentz transformations. In terms of the so-called chiral representation (2.65), the projectors are found to take the following explicit form

$$\mathcal{P}_L = \frac{I - \gamma^5}{2} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \quad \text{and} \quad \mathcal{P}_R = \frac{I + \gamma^5}{2} = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}. \quad (2.144)$$

In this form, the projection properties of these operators become transparent. Note by the way that due to the fact that γ^5 and γ^0 anticommute (2.143), the joint spinors $\psi_{L,R}$ are defined using opposite projectors

⁵The operator γ^5 is called chirality because it indicates how its eigenstates transform under a parity transformation [6, 9].

$$\bar{\psi}_L = \bar{\psi} \left(\frac{I + \gamma^5}{2} \right) = \bar{\psi} \mathcal{P}_R \quad \text{and} \quad \bar{\psi}_R = \bar{\psi} \left(\frac{I - \gamma^5}{2} \right) = \bar{\psi} \mathcal{P}_L. \quad (2.145)$$

Thus, they satisfy

$$\bar{\psi} \gamma^\mu \psi = \bar{\psi}_L \gamma^\mu \psi_L + \bar{\psi}_R \gamma^\mu \psi_R \quad (2.146)$$

and

$$\bar{\psi} \psi = \bar{\psi}_L \psi_R + \bar{\psi}_R \psi_L, \quad \text{and} \quad \bar{\psi}_L \psi_L = \bar{\psi}_R \psi_R = 0, \quad (2.147)$$

which confirms that a spinor existing only in one chirality state, cannot have any mass term in the Lagrangian.

From the projectors (2.144), one can deduce that the eigenstates of the chirality are quadri-spinors

$$\psi_L = \begin{pmatrix} \xi \\ 0 \end{pmatrix} \quad \text{and} \quad \psi_R = \begin{pmatrix} 0 \\ \chi \end{pmatrix},$$

where ξ and χ are two bi-spinors representing respectively the left and right components of the total spinor. The Dirac equation (2.62) can then be written as two coupled equations, the Weyl equations, for the bi-spinors

$$(\partial_t + \sigma \cdot \nabla) \xi = -im\chi, \quad (2.148)$$

$$(\partial_t - \sigma \cdot \nabla) \chi = -im\xi, \quad (2.149)$$

where the mass m should now be understood as a coupling term between the two components. For a massless spinor, both equations decouple and each bi-spinor, ξ and χ , called Weyl spinors, evolve independently. We will see later that these kinds of spinors have an especially important role to play in the formalism of supersymmetry.

2.6.2.3 Leptons

For a long time, neutrinos were thought to be represented by some Weyl tensors. Experimentally, they are only observed in the left state ν_L , leading to the conclusion that they should be massless. Since they only have two degrees of freedom, in order to put them in similar doublets as in (2.141), it is only possible to pair them with the left component of the corresponding lepton, leaving the right component in a singlet of $SU(2)_L$, thus the index 'L' of the group. The structure is thus

$$\begin{pmatrix} \nu_{e_L} \\ e_L \end{pmatrix}, \quad \begin{pmatrix} \nu_{\mu_L} \\ \mu_L \end{pmatrix}, \quad \begin{pmatrix} \nu_{\tau_L} \\ \tau_L \end{pmatrix}, \quad (2.150)$$

which gathers in a similar scheme, the electron, the muon and the tau with their associated neutrinos.

The covariant derivatives are defined as

$$D_\mu^{(2)} \begin{pmatrix} \nu_{e_L} \\ e_L \end{pmatrix} \equiv (\partial_\mu - ig_2 B_{i\mu} \sigma^i) \begin{pmatrix} \nu_{e_L} \\ e_L \end{pmatrix}. \quad (2.151)$$

σ^i ($i = 1, 2, 3$) are the Pauli matrices (2.66), generators of the group $SU(2)$ since they satisfy the algebra

$$[\sigma^i, \sigma^j] = 2i\epsilon^{ij}_k \sigma^k. \quad (2.152)$$

Similar to the strong interaction, the vector bosons $B_{i\mu}$ are in the adjoint representation of $SU(2)$. The coupling constant of the group is here called g_2 . The coefficients



ϵ^{ij}_k take numerical values +1 (resp. -1) if (i, j, k) is an even (resp. odd) permutation of $(1, 2, 3)$, and 0 otherwise (i.e. if two indices are the same). The structure constants of the group $SU(2)$ are given by the rank 3 Levi-Civita tensor⁶, and this group is locally equivalent to that of rotations.

Moreover, a charge with respect to the phase transformations is associated to each particle. This ‘hypercharge’, to distinguish it from the electric charge Q , is denoted by Y and we have $Y_R = -2$ for the right leptons, $Y_L = -1$ for the left leptons, $Y(q_L) = \frac{1}{3}$ for the quark doublets (2.141), of left chirality, and for the right chirality, $Y(u_R, c_R, t_R) = \frac{4}{3}$ and $Y(d_R, s_R, b_R) = -\frac{2}{3}$.

The doublet structure of (2.141) and (2.150) reminds us of that of spin $\frac{1}{2}$ with its components $|\uparrow\rangle$ and $|\downarrow\rangle$. This is why the doublet classification of $SU(2)_c$ is also called weak isospin, denoted by T to avoid any confusion with the spin S . The values $T^3 = \pm \frac{1}{2}$ are respectively given to the top and bottom component of the doublet. Using this analogy and bearing in mind that the electric charges of the quarks are $Q_{u,c,t} = +\frac{2}{3}$ and $Q_{d,s,b} = -\frac{1}{3}$, the ones for the massive leptons $Q_{e,\mu,\tau} = -1$ and that neutrinos have no electric charge, we obtain the Gell-Mann–Nishijima relation

$$Q = T^3 + \frac{Y}{2}. \quad (2.153)$$

Note that for this relation to hold for all particles, we need to set $T = 0$ for the singlets. We can then add a covariant derivative term $U(1)_Y$ for each field,

$$D^{(1)}\psi \equiv (\partial_\mu + ig_1 Y C_\mu)\psi,$$

where g_1 is the coupling constant associated with the group $U(1)_Y$.

2.6.3 A complete model

Putting together all the previous building blocks, a complete Lagrangian can be written down to describe the three non-gravitational interactions and all their invariances.

2.6.3.1 Kinetic terms

The kinetic terms of the standard model for the electroweak and strong interactions are thus given by

$$\begin{aligned} \mathcal{L}^{\text{kin}} = & -i\bar{\psi}\gamma^\mu(\partial_\mu - ig_3 G_{a\mu}\lambda^a - ig_2 B_{i\mu}\sigma^i - ig_1 Y C_\mu)\psi \\ & -\frac{1}{4}(\partial_\mu C_\nu - \partial_\nu C_\mu)^2 \\ & -\frac{1}{4}(\partial_\mu G_{a\nu} - \partial_\nu G_{a\mu} + if^{bc}_a G_{b\mu}G_{c\nu})^2 \\ & -\frac{1}{4}(\partial_\mu B_{i\nu} - \partial_\nu B_{i\mu} + i\epsilon^{jk}_i B_{j\mu}B_{k\nu})^2, \end{aligned} \quad (2.154)$$

⁶The Levi-Civita tensor ϵ in n dimensions is the completely antisymmetric rank n tensor with components $\epsilon_{1,\dots,n} = +1$

where the action of the covariant derivative depends on which particle it acts. For instance, for the lepton doublet of the first family, using the hypercharges obtained earlier, this gives,

$$\mathcal{L}_e^{\text{kin}} = -i(\bar{\nu}_{e_L}, \bar{e}_L) \gamma^\mu (\partial_\mu - ig_2 B_{i\mu} \sigma^i + ig_1 C_\mu) \begin{pmatrix} \nu_{e_L} \\ e_L \end{pmatrix} - i\bar{e}_R \gamma^\mu (\partial_\mu + 2ig_1 C_\mu) e_R, \quad (2.155)$$

since the leptons are not subject to the strong interaction.

2.6.3.2 W^\pm and Z^0 bosons and photon; Weak angle

Equation (2.155) can explicitly be expanded in left and right components as

$$\begin{aligned} \mathcal{L}_e^{\text{kin}} = & -i\bar{\nu}_{e_L} \gamma^\mu \partial_\mu \nu_{e_L} - i\bar{e} \gamma^\mu (\partial_\mu + 2ig_2 \sin \theta_w A_\mu) e \\ & - \sqrt{2}g_2 (\bar{\nu}_{e_L} \gamma^\mu W_\mu^+ e_R + \bar{e}_R \gamma^\mu W_\mu^- \nu_{e_L}) \\ & - \frac{g_2}{\cos \theta_w} (\bar{\nu}_{e_L} \gamma^\mu Z_\mu \nu_{e_L} - \cos 2\theta_w \bar{\nu}_{e_L} \gamma^\mu Z_\mu e_L - 2 \sin^2 \theta_w \bar{\nu}_{e_R} \gamma^\mu Z_\mu e_R), \end{aligned} \quad (2.156)$$

where the weak angle θ_w is defined by the relations

$$\left. \begin{aligned} \sin \theta_w &\equiv \frac{g_1}{\sqrt{g_1^2 + g_2^2}} \\ \cos \theta_w &\equiv \frac{g_2}{\sqrt{g_1^2 + g_2^2}} \end{aligned} \right\} \implies \frac{g_1}{g_2} = \tan \theta_w. \quad (2.157)$$

As for the vector fields, they are defined by

$$W_\mu^\pm = \frac{1}{\sqrt{2}} (B_{1\mu} \mp iB_{2\mu}), \quad (2.158)$$

and

$$\begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix} \equiv \begin{pmatrix} \cos \theta_w & -\sin \theta_w \\ \sin \theta_w & \cos \theta_w \end{pmatrix} \begin{pmatrix} B_{3\mu} \\ C_\mu \end{pmatrix} \iff \begin{pmatrix} B_{3\mu} \\ C_\mu \end{pmatrix} \equiv \begin{pmatrix} \cos \theta_w & \sin \theta_w \\ -\sin \theta_w & \cos \theta_w \end{pmatrix} \begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix}, \quad (2.159)$$

A_μ being the electromagnetic field.

The Lagrangian (2.156) can be understood in the following way. First, we notice that the kinetic term of the electron leads to a gauge coupling of the $U(1)$ type with the photon, since it can be written in the form

$$i\bar{e} \gamma^\mu D_\mu^{\text{elec}} e \equiv i\bar{e} \gamma^\mu (\partial_\mu - 2iqQ A_\mu) e,$$

with

$$q = g_2 \sin \theta_w$$

the electromagnetic coupling constant and Q the electric charge operator, with value $Q_e = -1$ for the electron. This coupling is the same for the left and right degrees of

freedom of the electron. In addition to the coupling terms, an intermediate neutral boson Z_μ has appeared, coupling both right and left components of the electron in a different way, and also adding a self-coupling term for the (left-handed) neutrino. Finally, charged intermediate bosons W^\pm make interactions between neutrinos and electrons possible (the charges of the W^\pm come from the fact that each term of the final Lagrangian should preserve the electric charge, which between a neutral neutrino and an electron is only possible with the indicated signs for the charges).

For a quark doublet, one can perform the same expansion, which gives us the electric charges of the quarks, but leads to 36 different interaction terms for each generation of quarks once the triplet of colour is made explicit.

2.6.3.3 Symmetry breaking

The Lagrangian (2.154) counts almost all possible terms that satisfy the invariance of the standard model and bringing into play the particles known experimentally. Due to its invariances, and in particular due to the weak isospin, it is not possible to write mass terms for the leptons. As for the quarks, all elements of the same doublet of $SU(2)_L$ should have the same mass, like for instance both quarks u and d , which is in contradiction with the measurements. Thus the model can only be made compatible with experiments if the $SU(2)_L$ symmetry is broken.

For this, we introduce a complex field doublet

$$\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix},$$

the dynamics of which is governed by the potential (2.128)

$$V(\phi) = \lambda (|\phi|^2 - \eta^2)^2 = \lambda (\phi_1^* \phi_1 + \phi_2^* \phi_2 - \eta^2)^2.$$

The minimum of this potential is $|\phi_1|^2 + |\phi_2|^2 = \eta^2$. Assuming that the Higgs field is not subject to the strong interaction, the kinetic term becomes

$$\mathcal{L}_\phi^{\text{kin}} = -|D\phi|^2 = -\left|(\partial_\mu - ig_2 B_{i\mu} \sigma^i - ig_1 Y_\phi C_\mu) \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}\right|^2, \quad (2.160)$$

which can be expanded as

$$\begin{aligned} |D\phi|^2 &= |\partial\phi_1|^2 + |\partial\phi_2|^2 + \sqrt{2}g_2 (\mathcal{J}_\mu^- W^{+\mu} + \mathcal{J}_\mu^+ W^{-\mu}) \\ &\quad + (g_2 B_{3\mu} + g_1 Y_\phi C_\mu) \mathcal{J}_1^\mu - (g_2 B_{3\mu} - g_1 Y_\phi C_\mu) \mathcal{J}_2^\mu \\ &\quad + (g_2 B_{3\mu} + g_1 Y_\phi C_\mu)^2 |\phi_1|^2 + (g_2 B_{3\mu} - g_1 Y_\phi C_\mu)^2 |\phi_2|^2 \\ &\quad + \sqrt{2}g_2 (\phi_1^* \phi_2 + \phi_1 \phi_2^*) [W^{-\mu} (g_2 B_{3\mu} + g_1 Y_\phi C_\mu) - W^{+\mu} (g_2 B_{3\mu} - g_1 Y_\phi C_\mu)] \\ &\quad + 2g_2 (|\phi_1|^2 + |\phi_2|^2) W_\mu^+ W^{-\mu}. \end{aligned} \quad (2.161)$$

In (2.160), we have introduced the $U_Y(1)$ hypercharge Y_ϕ of the Higgs doublet; we determine its value below.

The currents are defined by

$$\mathcal{J}_{1,2}^\mu \equiv i(\phi_{1,2}^* \partial^\mu \phi_{1,2} - \phi_{1,2} \partial^\mu \phi_{1,2}^*),$$

and

$$\mathcal{J}_\mu^- \equiv i(\phi_1^* \partial_\mu \phi_2 - \phi_2 \partial_\mu \phi_1^*),$$

with $\mathcal{J}^+ = (\mathcal{J}^-)^\dagger = i(\phi_2^* \partial_\mu \phi_1 - \phi_1 \partial_\mu \phi_2^*)$.

We now work around the minimum of the potential, and perform a gauge transformation to suppress any of the three redundant degrees of freedom. In the unitary gauge, for which $\phi_1 = 0$, and $\phi_2 = \eta + h/\sqrt{2}$, i.e.

$$\phi = \phi_0 + \delta\phi, \quad \text{with} \quad \phi_0 = \begin{pmatrix} 0 \\ \eta \end{pmatrix} \quad \text{and} \quad \delta\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ h \end{pmatrix}, \quad (2.162)$$

we are left with only the Higgs field, h , which is real. We then get

$$\mathcal{L}_\phi^{\text{kin}} = -\frac{1}{2} \partial_\mu h \partial^\mu h - \left(\eta^2 + \frac{1}{2} h^2 + \sqrt{2}\eta h \right) \left[(g_2 B_{3\mu} - g_1 Y_\phi C_\mu)^2 + 2g_2^2 W^{+\mu} W_\mu^- \right], \quad (2.163)$$

where we can recognize the intermediate vector boson Z_μ provided that we fix the hypercharge of the Higgs scalar doublet to $Y_\phi = 1$. It is the only value that will preserve the electromagnetic invariance after the symmetry breaking, ensuring the photon defined by the relation (2.159) remains massless.

2.6.3.4 Masses of the intermediate bosons

Equation (2.163) contains interaction terms between the Higgs boson, h , and the intermediate bosons, Z^0 and W^\pm . The mass term of the latter can be extracted. Taking $h = 0$ in (2.163), we get

$$\mathcal{L}_{\text{masses}} = -2g_2^2 \eta^2 W_\mu^+ W^{-\mu} - \eta^2 \frac{g^2}{\cos^2 \theta_w} Z_\mu Z^\mu \equiv -M_W^2 W_\mu^+ W^{-\mu} - \frac{1}{2} M_Z^2 Z_\mu Z^\mu, \quad (2.164)$$

so that we can make the identification

$$M_W = \sqrt{2}g_2\eta \quad \text{and} \quad M_Z = \frac{\sqrt{2}g_2\eta}{\cos \theta_w} = \frac{M_W}{\cos \theta_w}. \quad (2.165)$$

This last relation allows us to derive the value of the weak angle θ_w from the masses of the intermediate bosons. This has been done following their discovery at CERN in 1982. Their actual masses are [11]

$$M_W = 80.419 \pm 0.056 \text{ GeV} \quad \text{and} \quad M_Z = 91.1882 \pm 0.0022 \text{ GeV},$$

which gives the weak mixing angle, namely $\sin^2 \theta_w \sim 0.22$, that is approximatively $\theta_w \simeq 28^\circ$. Note that no term of the type $A_\mu A^\mu$ appears in the Lagrangian (2.164). The photon remains massless, in agreement with Maxwell's electromagnetism.

2.6.3.5 Masses of the fermions

The particles of right and left chirality belong to different representations of $SU_c(2)$, thus it is not possible to construct mass terms in a simple way just as for the gauge fields, one should impose that all the fermions are massless before the symmetry breaking. Moreover, since the right neutrino should be absent from this theory, one should postulate that the neutrino is massless, even after symmetry breaking.

The most general coupling terms between the fermions and the Higgs field that satisfy the required invariances, are given by⁷

$$\mathcal{L}_{\text{Yukawa}} = - \sum_{\text{generations}} \left(f_{\text{leptons}} \bar{\ell}_L \cdot \phi e_R + f_{q\uparrow} \bar{q}_L \cdot \phi u_R + f_{q\downarrow} \bar{q}_L \cdot \phi d_R + \text{h.c.} \right), \quad (2.166)$$

with

$$\tilde{\phi} \equiv i\sigma^2 \phi^* = \begin{pmatrix} \phi_2^* \\ -\phi_1^* \end{pmatrix},$$

where the matrix σ^2 is defined by (2.66), with hypercharge $Y_{\tilde{\phi}} = -Y_\phi = -1$, and where the sum is over the three different generations, so that the fields e_R , u_R and d_R represent, respectively, $(e, \mu, \tau)_R$, $(u, c, t)_R$ and $(d, s, b)_R$. In (2.166), the numbers f_{leptons} , $f_{q\uparrow}$ and $f_{q\downarrow}$, called the Yukawa coefficients, not only represent the coupling of the quarks and leptons with the Higgs fields, but, once the symmetry is broken, also give the values of the masses of the different fermions. One can see that the interaction of a given particle with the Higgs field is all the more important that it is massive.

2.6.3.6 Electromagnetic symmetry

Working in the unitary gauge defined by (2.162), we see that the vacuum configuration ϕ_0 is identically annihilated by the electric charge operator Q given by the Gell-Mann-Nishijima relation (2.153). Indeed, with

$$T^3 = \frac{1}{2} \sigma^3 = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

and $Y_\phi = 1$, we have

$$Q\phi_0 = \left(T^3 + \frac{1}{2} \right) \phi_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \eta \end{pmatrix} = 0.$$

The annihilation of the vacuum state by the generator of electromagnetism implies that this state satisfies

$$e^{i\alpha(x^\mu)Q} \phi_0 = \phi_0, \quad (2.167)$$

and hence it is invariant under a transformation $U(1)_{\text{elec}}$. Thus, the symmetry breaking is not total and the electromagnetic invariance is recovered, as required for the theory to be compatible with experiments.

⁷ One can explicitly check that with the hypercharges given previously for all particles, these terms are the only possible ones that are invariants under the transformation of $U(1)_Y$.

Note that even if here the relation (2.167) has been explicitly obtained in a unitary gauge, from the invariance of the initial gauge of the theory we can conclude, independently of the gauge, that this result will apply for any choice of gauge, i.e. that for any choice of the configuration of the Higgs doublet, as long as it is at the minimum of the potential.

2.7 Discrete invariances

In addition to the continuous symmetries, usually gauged ones, there exist discrete symmetries such as parity, charge conjugation and time-reversal invariance. The first one appears to be maximally violated by the weak interactions, and the last one weakly, for instance, in the system $K_0 - \bar{K}_0$. In what follows, we will limit ourselves to the study of the scalar field, all results being generalizable to the fermionic and vector sectors [8].

For a classical field, a symmetry operation can be obtained in the following way. For a given transformation $x \mapsto f(x, \alpha)$, where α represents a set of parameters defining the transformation, the field $\phi(x)$ transforms as

$$\phi'(x') = A(\alpha)\phi(x), \quad (2.168)$$

where A is a matrix if the scalar field ϕ has several components.

A state $|\psi\rangle$ in the Fock space, corresponding to the field operator $\hat{\phi}$, transforms linearly $|\psi'\rangle = \hat{U}(\alpha)|\psi\rangle$, where $\hat{U}^\dagger = \hat{U}^{-1}$, ensuring that $\langle\psi|\psi\rangle = \langle\psi'|\psi'\rangle$. We require that the matrix elements of the field operator satisfy the relation (2.168),

$$\langle\psi'_1|\hat{\phi}(x')|\psi'_2\rangle = A(\alpha)\langle\psi_1|\hat{\phi}(x')|\psi_2\rangle$$

between two arbitrary states $|\psi_1\rangle$ and $|\psi_2\rangle$ and the transformed states $|\psi'_1\rangle$ and $|\psi'_2\rangle$, which leads to

$$\hat{U}^\dagger(\alpha)\hat{\phi}(x')\hat{U}(\alpha) = A(\alpha)\hat{\phi},$$

i.e. to

$$\hat{U}^{-1}(\alpha)\hat{\phi}(x')\hat{U}(\alpha) = A(\alpha)\hat{\phi}[f^{-1}(x, \alpha)],$$

where the dummy variable x' has been replaced by x and expressed using the inverse transformation law.

We can now see how these general transformations can be applied to different particular cases, useful in particle physics.

2.7.1 Parity

Let us assume that we perform a reflection transformation on the spatial coordinates

$$\begin{cases} x \mapsto x' = -x, \\ t \mapsto t' = t. \end{cases} \quad (2.169)$$

If the physics remains unchanged under this transformation (as is the case, for instance, for the Klein–Gordon equation that is of second order in the spatial derivatives), then a classical field can only be modified by at most a phase factor

$$\phi'(x') = s_p \phi(x),$$

with $|s_p|^2 = 1$. For the field operator this translates into

$$\hat{P}^{-1} \hat{\phi}(x, t) \hat{P} = s_p \hat{\phi}(-x, t), \quad (2.170)$$

where \hat{P} is a unitary operator representing the parity.

If $\phi \in \mathbb{C}$, then the variation can be performed via a phase, i.e. $s_p = e^{i\alpha}$, $\alpha \in \mathbb{R}$. However, for a real field ($\phi \in \mathbb{R}$), the corresponding operator should be Hermitian, i.e. $\hat{\phi}^\dagger = \hat{\phi}$, and hence

$$\hat{P}^\dagger \hat{\phi}^\dagger(x, t) (\hat{P}^{-1})^\dagger = \hat{P}^{-1} \hat{\phi}(x, t) \hat{P} = s_p^* \hat{\phi}(-x, t),$$

from which we deduce that $s_p^* = s_p$. This, together with the relation $|s_p|^2 = 1$, implies that $s_p = \pm 1$. In this special case, the particle associated with the field will be said to be scalar if $s_p = +1$ and pseudo-scalar for $s_p = -1$, i.e. when it has a negative intrinsic parity.

The action of the parity operator on the Hilbert space can be obtained by expanding the field in the basis (2.106) (we consider from now on a complex field) and rewriting the equation. We find

$$\begin{aligned} & \int \frac{d^3 k}{(2\pi)^{3/2} \sqrt{2\omega_k}} [\hat{P} \hat{a}_k \hat{P}^\dagger e^{i(k \cdot x - \omega_k t)} + \hat{P} \hat{b}_k^\dagger \hat{P}^\dagger e^{-i(k \cdot x - \omega_k t)}] \\ &= s_p \int \frac{d^3 k}{(2\pi)^{3/2} \sqrt{2\omega_k}} [\hat{a}_k e^{-i(k \cdot x + \omega_k t)} + \hat{b}_k^\dagger e^{i(k \cdot x - \omega_k t)}], \end{aligned} \quad (2.171)$$

where the sign of x in the plane-wave exponentials of the right-hand side have been inverted, in accordance with (2.170). Since this last operation is equivalent, in the exponentials, to the one that changes the sign of k , it gives the transformation laws of the creation and annihilation operators

$$\hat{P} \hat{a}_k \hat{P}^\dagger = s_p \hat{a}_{-k} \quad \text{and} \quad \hat{P} \hat{b}_k^\dagger \hat{P}^\dagger = s_p \hat{b}_{-k}^\dagger. \quad (2.172)$$

Thus, the parity operator reverses the sign of the momenta, which had to be expected since they are 3-vectors.

For the case of a free complex scalar field, i.e. massive but with no interaction with any other field, the Hamiltonian, given by the relation (2.108), transforms under parity as

$$\hat{P} \hat{H} \hat{P}^{-1} = |s_p|^2 \int d^3 k \omega_k [\hat{N}_{-k}^{(a)} + \hat{N}_{-k}^{(b)}] = \hat{H}, \quad (2.173)$$

since the integral is performed on all values of the vector k and $|s_p|^2 = 1$ whichever is the parity of the states we consider. Similarly, the field momentum, equal to

$$\hat{P} = \int d^3 k k [\hat{N}_k^{(a)} + \hat{N}_k^{(b)}], \quad (2.174)$$

transforms as

$$\hat{P} \hat{P} \hat{P}^{-1} = -\hat{P},$$

which corresponds to the transformation law of a vector. At this point, it is interesting to note from (2.173) that the parity operator commutes with the Hamiltonian, and thus

the parity of a physical state is a constant of motion for the free complex scalar field. Some interactions can modify this conclusion: this is the case of the weak interactions.

A parity can also be attributed to the vacuum, and we choose $\hat{P}|0\rangle = |0\rangle$ (the vacuum must remain identical to itself if the coordinates are inverted). Note that we can explicitly construct an operator \hat{P} satisfying (2.172) for which this relation is verified, but it is also possible to give a negative intrinsic parity to the vacuum: this only complicates things a little and simply inverts all the fields parities. In other words, this is only a matter of convention.

2.7.2 Charge conjugation

Another discrete transformation as simple as parity can also be performed. This is charge conjugation \hat{C} , which consists in replacing each particle in a given physical state by its antiparticle. In terms of operators, this amounts to transforming ϕ into ϕ^\dagger and vice versa. We can therefore write

$$\hat{C}\hat{\phi}\hat{C}^\dagger = s_c\hat{\phi}^\dagger \quad \text{and} \quad \hat{C}\hat{\phi}^\dagger\hat{C}^\dagger = s_c^*\hat{\phi}, \quad (2.175)$$

the second relation of (2.175), being simply the Hermitian conjugate of the first one. Just as with parity, in (2.175), the eigenvalue of s_c has a unit modulus since transforming twice in a row all the particles and their antiparticles is in the end equivalent to no transformation at all. This is equivalent to saying that $\hat{C}^2 = 1$. Furthermore, note that \hat{C} should be unitary, so that finally, we have $\hat{C}^\dagger = \hat{C}^{-1} = \hat{C}$.

In terms of the annihilation operators, we find

$$\hat{C}\hat{a}_k\hat{C}^\dagger = s_c\hat{b}_k \quad \text{and} \quad \hat{C}\hat{b}_k\hat{C}^\dagger = s_c^*\hat{a}_k, \quad (2.176)$$

and the Hermitian conjugate relations for the creation operators. Assuming, as before, that the vacuum is an eigenstate of the charge conjugation, with eigenvalue +1, i.e. $\hat{C}|0\rangle = |0\rangle$, we can then determine the charge \hat{C} of any physical state created with a determined number of particles. Here again, we find that for a free scalar field, the charge conjugation is a constant of motion, leaving the Hamiltonian (as well as the momentum) unchanged. However, the conserved charge \hat{Q} of (2.113) changes sign, as one could expect,

$$\hat{C}\hat{Q}\hat{C}^{-1} = -\hat{Q},$$

so that an eigenstate of the conserved charge \hat{Q} cannot simultaneously be an eigenstate of the charge conjugation. The time conservation of the latter is thus only useful for physical states with a global vanishing charge.

Most of the interactions satisfy $[\hat{H}, \hat{C}] = 0$, and the ones that violate parity conservation, such as the weak interactions, more often than not, leave the combination $\hat{C}\hat{P}$ invariant. In the section dedicated to baryogenesis in Chapter 9, we will discuss the cases where even this constraint is not valid. However, in all field theories analogous to the one presented here, a discrete symmetry is satisfied, it is the combination $\hat{C}\hat{P}\hat{T}$, where the last symmetry, the time-reversal invariance, requires some technical details that are given in what follows.

2.7.3 Time reversal and CPT theorem

The last discrete invariance, valid for instance for a free scalar field, is that related to time reversal \hat{T} :

$$\begin{cases} x \mapsto x' = x, \\ t \mapsto t' = -t. \end{cases} \quad (2.177)$$

Unlike the two previous symmetries, which are unitary operators, this symmetry can be implemented in the Fock space of states only in the form of an antiunitary operator.⁸ An operator \hat{T} is antiunitary if it is antilinear, i.e. such that

$$\hat{T}(\alpha|\psi\rangle + \beta|\xi\rangle) = \alpha^* \hat{T}|\psi\rangle + \beta^* \hat{T}|\xi\rangle, \quad (2.178)$$

and if it satisfies $\hat{T}\hat{T}^\dagger = \hat{T}^\dagger\hat{T} = 1$.

Thanks to this last operator, one can prove [14] the following very general theorem. If a local quantum field theory can be represented by a Hermitian Lagrangian $\hat{\mathcal{L}}(x^\alpha)$ that is invariant under proper Lorentz transformations (i.e. without the space and time inversions) and if its field operators satisfy the spin-statistics theorem,⁹ then we have

$(\hat{C}\hat{P}\hat{T})\hat{\mathcal{L}}(x^\alpha)(\hat{C}\hat{P}\hat{T})^{-1} = \hat{\mathcal{L}}(-x^\alpha).$

(2.179)

Due to this relation (2.179), the action integral $\hat{S} = \int d^4x \hat{\mathcal{L}}$ is invariant under the transformation $\hat{C}\hat{P}\hat{T}$. Thus, the equations of motion, which are obtained by varying the action, and the canonical commutation relations, are also unchanged under this transformation. We can also deduce that the Hamiltonian of the given theory must commute with $\hat{C}\hat{P}\hat{T}$,

$$[\hat{C}\hat{P}\hat{T}, \hat{H}] = 0,$$

and hence any realistic theory, i.e. satisfying the hypothesis of this theorem, has this invariance. This result is often written in the symbolic form ' $\hat{C}\hat{P}\hat{T} = 1$ '.

A non-trivial consequence of this theorem is the following: for any theory satisfying it, i.e. for any realistic theory, the particles and antiparticles should have exactly the same masses and the same lifetimes. This consequence is verified with a very high precision since, for example, for the K_0/\bar{K}_0 system, we find [11] $|M_{K_0} - M_{\bar{K}_0}|/M_{K_0} < 10^{-18}$. For most of the other particles, we find typical values of order of $|\Delta M|/M < 10^{-5}$, and constraints on the differences of lifetimes of order of $|\Delta\tau|/\tau < 10^{-3}$.

⁸This happens because of technical reasons related to the fact that the time-reversal operator must also reverse the roles of the initial and final configurations [8].

⁹The spin-statistics theorem, itself also very general in field theory, specifies that the particles associated to integer-spin fields (for instance bosons, scalars and vectors) have a statistics described by the Bose-Einstein distribution function, whereas the ones corresponding to half-integer fields (fermions, Dirac spinor, for instance), are described by Fermi-Dirac statistics. This theorem only relies on microcausality. It is discussed in detail in Ref. [15].

The standard model of particle physics presented in this chapter has so far never displayed any flaws from the point of view of experimental physics. However no one sincerely thinks that it describes all interactions between the elementary particles in a final way. There are many reasons for this, which are as much experimental as theoretical.

From an experimental point of view, the recent discovery of the existence of a mass for the neutrinos and the discussion concerning the chirality (Section 2.6.2) indicate that they cannot uniquely exist in the left chirality as attributed by the standard model. There must therefore exist a right neutrino, which has different properties from the left one since it would have otherwise already been detected experimentally. They therefore need to have a very large mass, in order to explain why no accelerator has already produced it.

At the experimental level, there is still an unknown. Since the discovery of the 'top' quark, all particles of the standard model are identified except one, which is maybe the most important since the principle of symmetry breaking relies on it, namely the Higgs boson. As long as it has not been produced directly in an accelerator, it will remain hypothetical, and the model will not have been fully verified.

Now from a theoretical point of view, the situation is even worse. First, this model, although it unifies in a quantum framework the three non-gravitational interactions (electromagnetism, weak and strong), makes no mention of the last interaction, gravity, despite the fact that one of its pillars, the Higgs mechanism, is supposed to be at the origin of the mass of elementary particles. Interestingly, these masses are also the gravitational charges of these particles. Moreover, the standard model relies on three different symmetry groups, with three coupling constants having no links between them, which is not justified by any fundamental principle. Finally, and this may be the most serious, there are 19 free parameters in the theory, which can only be measured. This proliferation of parameters is often considered as the sign that it is not a fundamental theory but actually a low-energy effective theory. We will see later that in many extensions, the number of free parameters is seriously reduced, and that it is in principle possible to calculate the 19 arbitrary coefficients that appear in the general Lagrangian in terms of a much smaller set of parameters.

References

- [1] L. D. LANDAU and E. LIFSHITZ, *Mechanics*, Pergamon Press, 1976.
- [2] N. DERUELLE and J. P. UZAN, *Mécanique et gravitation newtoniennes*, Vuibert, 2006.
- [3] C. COHEN-TANNOUDJI, B. DJU and F. LALOË, *Quantum mechanics*, John Wiley, 1992.
- [4] P. R. HOLLAND, *The quantum theory of motion*, Cambridge University Press, 1993.
- [5] J. SCHWINGER, *Quantum mechanics*, Springer, 2001.
- [6] F. HALZEN and A. D. MARTIN, *Quarks & leptons*, John Wiley and Sons, 1984.
- [7] M. A. PESKIN and D. V. SCHROEDER, *An introduction to quantum field theory*, Addison-Wesley, 1995.
- [8] W. GREINER and J. REINHARDT, *Field quantization*, Springer-Verlag, 1996.
- [9] L. H. RYDER, *Quantum field theory*, Cambridge University Press, 1987.
- [10] J. I. KAPUSTA, *Finite-temperature field theory*, Cambridge Monographs on Mathematical Physics, 1994
- [11] W.-M. YAO *et al.*, ‘The review of particle physics’, *Journal of Physics, G* **33**, 1, 2006; see also the website <http://pdg.lbl.gov/> that gives the same information. Updated almost every two years to take into account new experiments and their results.
- [12] E. SEGRÉ, *From X-Rays to quarks: modern physicists and their discoveries*, Freeman, 1980.
- [13] H. GEORGI, *Lie Algebras in particle physics*, Westview Press, 1999.
- [14] W. PAULI, in *Niels Bohr and the development of physics*, Pergamon Press, 1955; G. LÜDERS, *Ann. Phys. (NY)* **2**, 1, 1957.
- [15] B. DJU *et al.*, *Physique statistique*, Hermann, 1996.

Part II

The modern standard cosmological model

3

The homogeneous Universe

The first cosmological solution of Einstein's equation was given by Einstein himself in 1917. At the expense of introducing a cosmological constant, he was able to construct a closed and static space. The general cosmological solutions were discovered independently by Alexandre Friedmann and Georges Lemaître in 1922 and 1927.

This chapter details the derivation of these solutions from Einstein's equations and studies their kinematics and dynamics before defining the various observational quantities that will be used in this book. It ends on some considerations on less-symmetric space-times. For complementary developments, see Refs. [1–7].

3.1 The cosmological solution of Friedmann–Lemaître

3.1.1 Constructing a Universe model

Within the framework of general relativity, one can try to answer the question of which solution of Einstein's equation describes the Universe we observe, or more modestly, which solution is an idealized (but good) model of our Universe. To carry out this program, we need enough input from astrophysical and cosmological observations.

This program is indeed difficult to tackle and its solution will, in particular, depend on the amount of available data. But, as we noted in the introduction, even with ideal data, there are limitations to the answers one can give to this problem. These limitations arise mainly because we observe only a finite part of one Universe (it is a unique object and we cannot discuss its probable nature by comparing it to similar objects) from a given space-time position (and we are unable to choose this point).

Astrophysical data are mainly localized on our past light cone and around our past worldline for geophysical data. They are thus localized on (just a portion of) a 3-dimensional hypersurface. First, it may be that various space-times are compatible with the same data and second, the interpretation of these data is not independent of the space-time structure.

Thus, we must distinguish between the *observable Universe* (for which, by definition, we have data) and the *Universe*, which includes regions we cannot directly influence or experiment. The inference of the geometrical properties of the Universe from the observable Universe cannot be achieved without hypothesis and philosophical prejudices.

The standard cosmological model, in its simplest form, relies on four central hypotheses.

H1 *Gravitation is well described by general relativity.* Indeed, as discussed in Chapter 1, general relativity is well tested on small scales (e.g., Solar System scales),

but we cannot exclude that it is not the case on large scales or at early times (see, e.g., Chapter 12 for examples of theories that dynamically become more and more similar to general relativity during the evolution of the Universe). In particular, the Einstein equivalence principle implies that the laws of physics can be extrapolated at all times (see Chapter 1). We will thus keep applying, as long as possible, the standard laws of physics, in particular for non-gravitational interactions.

- H2 Hypothesis on the nature of matter.** We do not examine the space-time itself or its matter distribution. Rather, we observe particular objects (e.g., galaxies,...) that are supposed to be representative of the total matter content. The influence of the variations of the properties of individual objects on the inference of the matter content of the Universe has to be quantified. On large scales, we will assume that matter is described by a mixture of a pressureless fluid and radiation, allowing also for a cosmological constant. In particular, we assume that there is no matter outside the standard model of particle physics described in Chapter 2. The perfect fluid hypothesis will be shown to derive from the hypothesis H3.
- H3 Uniformity principles.** The previous hypotheses are not sufficient to solve the Einstein equations. We must assume some symmetries for the space-time. This hypothesis and their consequences are discussed in Section 3.1.2
- H4 Global structure.** Hypotheses H1–H3 enable us, as we shall see, to solve Einstein’s equation and to determine the local structure of the Universe model. However, there may exist various manifolds with identical local geometry and different global topology, which correspond to the choice of various boundary conditions. For a globally hyperbolic space-time, the choice of the topology reduces to the choice of the topology of the spatial sections.

As we shall see, these hypotheses allow us to construct very successful Universe models but there are reasons to challenge every one of these four assumptions. This will be discussed mainly in Chapter 12.

3.1.2 Cosmological and Copernican principles

Galaxies seem to be spread isotropically around us, just as the cosmic microwave background radiation. This indicates that the space-time describing the observable Universe should have a spherical symmetry around us. There are two possible explanations: either our Galaxy lies in a special place in the Universe so that it is not homogeneous and must have a centre. Or the space-time is homogeneous; every point is then similar and the Universe is isotropic around each point (see Fig. 3.1).

To construct cosmological models, we can distinguish two uniformity principles. The *cosmological principle* supposes that the Universe is spatially isotropic and homogeneous. In particular, this implies that any observer sees an isotropic Universe around him. We can distinguish it from the *Copernican principle* that merely states that we do not lie in a special place (the centre) of the Universe. This principle, together with the isotropic hypothesis, actually implies the cosmological principle.

The cosmological principle makes definite predictions about all unobservable regions beyond the observable Universe. It completely determines the entire structure of the Universe, even for regions that cannot be observed. From this point of view, this hypothesis, which cannot be tested, is very strong. The Copernican principle has

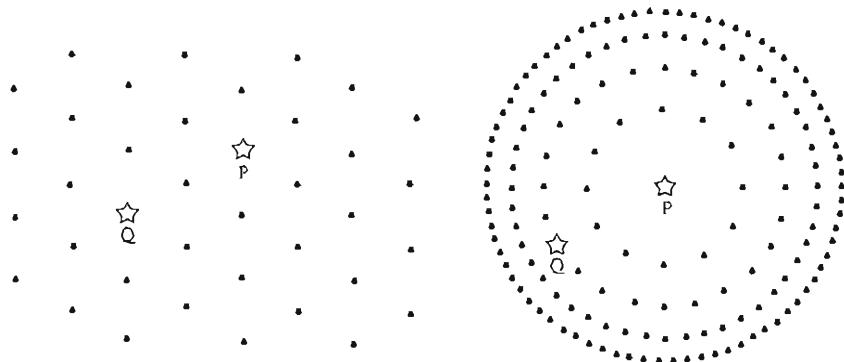


Fig. 3.1 A point distribution, statistically isotropic around every point (left) and around a unique point (P) (right). In the second version, P and Q are not equivalent. The cosmological principle excludes such kinds of solutions, which would assume that we lie in a special place in the Universe. From Ref. [1] of the introduction.

more modest consequences and leads to the same conclusions but only for the observable Universe where isotropy has been verified. It does not make any prediction on the structure of the Universe for unobserved regions (in particular, space could be homogeneous and non isotropic on scales larger than the observable Universe). Some cosmological spaces that do not satisfy this cosmological principle are described in Section 3.6.

From a pragmatic point of view, these uniformity principles should be considered in a statistical sense, i.e. for a Universe smoothed on distances greater than its larger structures. Thus, they implicitly involve a *smoothing scale* and a smoothing procedure to which we shall return. They lead to the Friedmann-Lemaître solutions that are at the basis of modern cosmology. Recently, a test of the Copernican principle was proposed [8].

3.1.3 Cosmological principle and metric

Homogeneity and isotropy are two distinct notions that we should define more precisely.

The *homogeneity* of space means that at every moment, each point of space is similar to any other one. More precisely, a space-time is (spatially) homogeneous if there exists a one-parameter family of space-like hypersurfaces, Σ_t , foliating the space-time such that for any t and for any pair of points (P, Q) of Σ_t , there exists an isometry of the space-time metric taking P into Q (see Fig. 3.2).

Isotropy means that at every point the Universe can be seen as isotropic. More precisely, a space-time is (spatially) isotropic at each point if there exists a congruence of time-like worldlines with tangent vector u^μ (called isotropic observers) such that at any point P and for any pairs of unit spatial tangent vectors $e_1^\mu, e_2^\mu \in V_P$ (see Chapter 1), there is an isometry of the metric that leaves P and u^μ at P fixed and that rotates the vectors e_1^μ into e_2^μ . Thus, it is impossible to construct a preferred tangent vector perpendicular to u^μ .

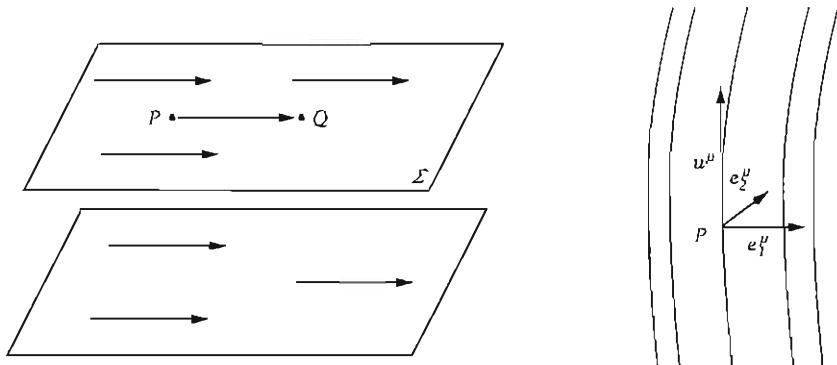


Fig. 3.2 (left): A homogeneous space-time can be foliated by a family of homogeneous spaces; (right): for an isotropic space, there is an isometry that rotates any unit spatial tangent vector e_1^μ to any other e_2^μ , while the vector u^μ remains invariant. From Ref. [9].

We recognize that for a homogeneous and isotropic space-time, the surfaces of homogeneity, Σ_t , must be orthogonal to the tangent vector u^μ of the worldline of the isotropic observers, that are thus called *fundamental observers*. (Proof: If it were not the case; then, assuming that the homogeneous space and the congruence of isotropic observers are unique, the failure of the tangent subspace perpendicular to u^μ to coincide with Σ_t would enable us to construct a geometrically privileged spatial direction. This is in contradiction with isotropy.)

The metric $g_{\mu\nu}$ induces a spatial metric $\hat{\gamma}_{\mu\nu}(t)$ on each hypersurface Σ_t . As we have just seen, there must exist an isometry of $\hat{\gamma}_{\mu\nu}$ bringing any point of Σ_t into another point of Σ_t and it must be impossible to construct a preferred vector from $\hat{\gamma}_{\mu\nu}$. As one can show, this implies that the Riemann tensor constructed from $\hat{\gamma}_{\mu\nu}$ for the three-dimensional spaces Σ_t must be of the form¹ [2, 9, 10]

$${}^{(3)}\hat{R}_{\alpha\beta\mu\nu} = 2\hat{K}\hat{\gamma}_{\mu[\alpha}\hat{\gamma}_{\beta]\nu}$$

where \hat{K} has, because of homogeneity, the same value at any point of Σ_t , and can only be a function of t , $\hat{K}(t)$. A space satisfying this property is called a *maximally symmetric space* (as we shall see it has the maximum number of Killing vectors allowed by its dimension) and is a space of constant curvature. Two spaces of constant curvature with the same dimension, signature and the same value of K are isometric. Thus, it is sufficient to enumerate the three-dimensional spaces corresponding to all values of \hat{K} .

We classify these spaces according to the sign of \hat{K} . If \hat{K} is positive, Σ_t is a three-dimensional sphere of radius $R(t)$, with embedding equation $x^2 + y^2 + z^2 + w^2 = R^2(t)$ in a four-dimensional Euclidean space. In spherical coordinates,

¹ As shown in Ref. [9], this conclusion can be drawn from the isotropy argument. Consider $(3)\hat{R}_{\mu\nu}{}^{\alpha\beta}$ as the components of a linear map, L , of the vector space of 2-forms into itself. Since L is symmetric, it can be diagonalized. If its eigenvalues were not equal, then one could construct a preferred 2-form and then a preferred direction. Thus, isotropy implies that L is proportional to the identity and thus $(3)\hat{R}_{\mu\nu}{}^{\alpha\beta} = 2\hat{K}\delta_{[\mu}^\alpha\delta_{\nu]}^\beta$ by symmetry.

$$x = R \cos \chi, \quad y = R \sin \chi \cos \theta, \quad z = R \sin \chi \sin \theta \cos \varphi, \quad w = R \sin \chi \sin \theta \sin \varphi,$$

the Euclidean metric $ds^2 = dx^2 + dy^2 + dz^2 + dw^2$ induces the three-dimensional metric on Σ_t

$$ds_{(3)}^2 = R^2(t) [d\chi^2 + \sin^2 \chi (d\theta^2 + \sin^2 \theta d\varphi^2)].$$

If $\hat{K} = 0$, Σ_t is a three-dimensional Euclidean space with metric

$$ds_{(3)}^2 = d\chi^2 + \chi^2 (d\theta^2 + \sin^2 \theta d\varphi^2).$$

If \hat{K} is negative, Σ_t is a three-dimensional hyperboloid, with embedding equation $-w^2 + x^2 + y^2 + z^2 = -R^2(t)$. In spherical coordinates,

$$x = R \cosh \chi, \quad y = R \sinh \chi \cos \theta, \quad z = R \sinh \chi \sin \theta \cos \varphi, \quad w = R \sinh \chi \sin \theta \sin \varphi,$$

the Minkowski metric $ds^2 = -dw^2 + dx^2 + dy^2 + dz^2$ induces the three-dimensional metric on Σ_t

$$ds_{(3)}^2 = R^2(t) [d\chi^2 + \sinh^2 \chi (d\theta^2 + \sin^2 \theta d\varphi^2)].$$

Although special relativity can only allow for one space with no curvature, general relativity offers three possibilities even under the strong constraint of homogeneity and isotropy.

3.1.3.1 Four-dimensional metric

The worldlines of the fundamental observers are orthogonal to the hypersurfaces Σ_t . The coordinate system can be transported from one hypersurface to another by means of this congruence of worldlines.

The general form of the space-time metric is imposed by these symmetry considerations

$$g_{\mu\nu} = -u_\mu u_\nu + \hat{\gamma}_{\mu\nu}(t),$$

where, for each value of t , $\hat{\gamma}_{\mu\nu}$ is the metric of a three-dimensional sphere, Euclidean space, or hyperboloid. It follows that the line-element of the cosmological space takes the form

$$ds^2 = -(u_\mu dx^\mu)^2 + \hat{\gamma}_{\mu\nu}(t)dx^\mu dx^\nu.$$

Setting $u_\mu dx^\mu \equiv dt$, the coordinate t is the proper time measured by the fundamental observers (i.e. comoving with u^μ) and we get the general form of the space-time metric as

$$ds^2 = -dt^2 + a^2(t)\gamma_{ij}(x^k)dx^i dx^j, \quad (3.1)$$

t being the cosmic time, $a(t)$ the scale factor and γ_{ij} the metric of constant time hypersurfaces, Σ_t , in comoving coordinates.² Latin indices i, j, \dots run from 1 to 3. The conformal time η , defined by

²This implies that $\hat{\gamma}_{ij}(x^k, t) = a^2 \gamma_{ij}(x^k)$, so that the spatial metric satisfies

$${}^{(3)}R_{ijkl} = 2K \gamma_{kl} \gamma_{ji}.$$

and $\hat{K}(t) = K/a^2(t)$.

$$dt = a(t)d\eta, \quad (3.2)$$

can also be introduced to recast the metric (3.1) in the form

$$ds^2 = a^2(\eta) (-d\eta^2 + \gamma_{ij}dx^i dx^j). \quad (3.3)$$

The spatial metric, in comoving spherical coordinates, takes the general form

$$d\sigma^2 = \gamma_{ij}dx^i dx^j = d\chi^2 + f_K^2(\chi)d\Omega^2, \quad (3.4)$$

where χ is a radial coordinate and $d\Omega^2 = d\theta^2 + \sin^2\theta d\varphi^2$ is the infinitesimal solid angle, φ runs from 0 to 2π and θ from $-\pi$ to π . The quantity $\sqrt{|K|}\chi$ runs from 0 to π when $K > 0$, and from 0 to $+\infty$ otherwise. The function f_K depends on the curvature K (which now is a pure number) as

$$f_K(\chi) = \begin{cases} K^{-1/2} \sin(\sqrt{K}\chi) & K > 0 \\ \chi & K = 0 \\ (-K)^{-1/2} \sinh(\sqrt{-K}\chi) & K < 0. \end{cases} \quad (3.5)$$

It relates the surface $S(\chi)$ of a comoving sphere to its radius χ by $S(\chi) = 4\pi f_K^2(\chi)$ that reduces to the Euclidean expression for distances small compared to the curvature (i.e. for $\chi \ll 1/\sqrt{|K|}$). Using the radial coordinate $r = f_K(\chi)$, (3.1) can also take the form

$$ds^2 = -dt^2 + a^2(t) \left(\frac{dr^2}{1 - Kr^2} + r^2 d\Omega^2 \right). \quad (3.6)$$

This metric, or the equivalent forms (3.1) or (3.3), is called the Friedmann–Lemaître metric.

This construction shows how the cosmological principle, and thus the symmetry assumptions that follow, has allowed us to reduce the ten arbitrary functions of the space-time metric to a single function of one variable, $a(t)$, and a pure number, K .

3.1.3.2 Geometrical quantities

The expression for the Christoffel symbols can be obtained, using the definitions of Chapter 1 and the metric (3.1). Their only non-vanishing components are

$$\Gamma_{ij}^0 = Ha^2\gamma_{ij}, \quad \Gamma_{j0}^i = H\delta_j^i, \quad \Gamma_{jk}^i = {}^{(3)}\Gamma_{jk}^i, \quad (3.7)$$

where the ${}^{(3)}\Gamma_{jk}^i$ are the Christoffel symbols of the spatial metric γ_{ij} . The quantity $H = \dot{a}/a$ is the Hubble parameter, and a dot represents the derivative with respect to the cosmic time t .

The explicit form of ${}^{(3)}\Gamma_{jk}^i$ is not needed to compute the component of the Ricci tensor because the spatial metric γ_{ij} satisfies ${}^{(3)}R_{ijkl} = K(\gamma_{ik}\gamma_{jl} - \gamma_{il}\gamma_{jk})$, so that

${}^{(3)}R_{ij} = 2K\gamma_{ij}$, and ${}^{(3)}R = 6K$. It follows that the only non-vanishing components of the Ricci tensor are³

$$R_{00} = -3\frac{\ddot{a}}{a}, \quad R_{ij} = \left(2H^2 + \frac{\ddot{a}}{a} + 2\frac{K}{a^2}\right)a^2\gamma_{ij}, \quad (3.8)$$

from which we deduce the expression for the scalar curvature

$$R = 6\left(H^2 + \frac{\ddot{a}}{a} + \frac{K}{a^2}\right), \quad (3.9)$$

and for the non-vanishing components of the Einstein tensor

$$G_{00} = 3\left(H^2 + \frac{K}{a^2}\right), \quad G_{ij} = -\left(H^2 + 2\frac{\ddot{a}}{a} + \frac{K}{a^2}\right)a^2\gamma_{ij}. \quad (3.10)$$

3.1.3.3 Killing vectors

As explained in Chapter 1, to each symmetry of a space-time is associated a Killing vector field, ξ , satisfying (1.58). We have seen that for a space-time of dimension n , the maximal number of independent Killing vectors, and thus of symmetries, is $n(n+1)/2$.

For the Friedmann-Lemaître space-times, the spatial sections are maximally symmetric so that we expect to obtain 6 Killing vectors. They can be explicitly determined by considering the Weinberg coordinates

$$x^1 = f_K(x) \cos \theta, \quad x^2 = f_K(x) \sin \theta \cos \varphi, \quad x^3 = f_K(x) \sin \theta \sin \varphi,$$

in terms of which the spatial metric, and its inverse, take the form

$$\gamma_{ij} = \delta_{ij} + \frac{K\delta_{im}\delta_{jn}x^m x^n}{1 - Kr^2}, \quad \gamma^{ij} = \delta^{ij} - Kx^i x^j, \quad (3.11)$$

with $r^2 \equiv \delta_{mn}x^m x^n$.

The Killing equation, $\nabla_\mu\xi_\nu + \nabla_\nu\xi_\mu = 0$, gives for the 00-component, $\xi_0 = 0$, so that ξ^0 is a function of the spatial coordinates only, $\xi^0 = F(x^k)$. Then, the 0i and ij components take the form $\xi^i = g^{ki}D_k\xi^0$ and $\gamma_{ik}D_j\xi^k + \gamma_{jk}D_i\xi^k + 2H\gamma_{ij}\xi^0 = 0$, where D_i is the covariant derivative associated to γ_{ij} . They can be combined to get

$$D_i D_j F + \dot{H}a^2\gamma_{ij}F = 0.$$

Only $\dot{H}a^2$ may possibly depend on time in this equation. If this is the case, $\dot{H}a^2$ is not a pure number, then this implies that $F = 0$. We thus have six solutions to the Killing equations:

³As an example, consider the computation of R_{ij} . It can be decomposed as $R_{ij} = \partial_0\Gamma_{ij}^0 - \partial_j\Gamma_{i0}^0 + \Gamma_{0\rho}^\rho\Gamma_{ij}^0 - \Gamma_{\beta j}^\alpha\Gamma_{i\alpha}^\beta + {}^{(3)}R_{ij}$, where all terms involving no $\mu = 0$ component have been gathered to identify the Ricci tensor of the spatial metric. We then use that ${}^{(3)}R_{ij} = 2K\gamma_{ij} = 2(K/a^2)g_{ij}$.

- three Killing vectors associated to spatial translations

$$P_{[r]}^\mu : \quad P_{[r]}^0 = 0, \quad P_{[r]}^i = \delta_r^i \sqrt{1 - Kr^2} \quad r = 1, \dots, 3, \quad (3.12)$$

- three Killing vectors associated to spatial rotations

$$R_{[rs]}^\mu : \quad R_{[rs]}^0 = 0, \quad R_{[rs]}^i = \delta^{ir} x^s - \delta^{is} x^r \quad r, s = 1, \dots, 3. \quad (3.13)$$

When $\dot{H}a^2$ does not depend on time, $\dot{H}a^2 = K$, as we shall see from the Friedmann equations. This occurs for the Minkowski, de Sitter and anti-de Sitter spaces. It follows that there are *four* additional solutions associated to time translation and Lorentz boosts. They are, respectively, given by

$$T^\mu : \quad T^0 = \sqrt{1 - Kr^2}; \quad T^k = -Hx^k \sqrt{1 - Kr^2} \quad (3.14)$$

and

$$L_{[r]}^\mu : \quad L_{[r]}^0 = x^r; \quad L_{[r]}^k = H \begin{cases} \left[\frac{1}{2} \delta^{kr} (r^2 - \eta^2) - x^k x^r \right] & \text{for } K = 0, \\ (K \delta^{kr} - x^k x^r) & \text{for } K = \pm 1, \end{cases} \quad (3.15)$$

in terms of the conformal time η .

Because of the expansion of the Universe, these four vectors are no longer Killing vectors of the Friedmann-Lemaître space-time. They now satisfy the equation $\nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu = (\nabla_\alpha \xi^\alpha) g_{\mu\nu}/2$ and are called *conformal Killing vectors*. The symmetries involving time (i.e. time translation and Lorentz boosts) are lost due to the fact that the cosmological space-times are not static but they are still conformal to a static space-time, a property related to the existence of these four conformal Killing vectors.

3.1.4 Kinematics

Without knowing the evolution law for the scale factor, $a(t)$, a few general consequences can be obtained simply from the form of the Friedmann-Lemaître metric.

3.1.4.1 Fundamental observers

Notice first of all that in the metric (3.1), the coordinates x^i are comoving. The worldline of a fundamental observer with trajectory orthogonal to the hypersurfaces Σ_t is given by $x^0 = t$ and $x^i = \text{constant}$ (one can check that this is indeed a time-like geodesic). As a consequence, this observer has a four-velocity

$$u^\mu = \delta_0^\mu. \quad (3.16)$$

As seen earlier, this metric can be re-expressed in the form

$$ds^2 = -(u_\mu dx^\mu)^2 + \hat{\gamma}_{\mu\nu} dx^\mu dx^\nu, \quad (3.17)$$

where $\hat{\gamma}_{\mu\nu} = g_{\mu\nu} + u_\mu u_\nu$ is the projector tensor introduced in Section 1.2.5. The expansion of the Universe thus allows a comoving observer to naturally decompose any vector P^μ into a time-like and space-like component as

$$P^\mu = -(P^\nu u_\nu)u^\mu + \hat{\gamma}_\nu^\mu P^\nu. \quad (3.18)$$

In particular, all tools introduced in sections 1.2.5 and 1.3.2 of the first chapter can be used.

3.1.4.2 Hubble law

Let us now consider two comoving observers with trajectories given by $x = x_1$ and $x = x_2$. In the physical space, their separation is given by

$$r_{12} = a(t)(x_1 - x_2).$$

It follows that

$$\dot{r}_{12} = H r_{12}. \quad (3.19)$$

The farther the observers are from each other, the greater is their relative velocity. This is what is usually called the Hubble law. Notice that H depends a priori on time and can only be approximated as a constant for objects whose distance is small in comparison to the Hubble radius. Note that this law is a special case of the generalized Hubble law (1.119) with $\Theta = 3H$ and $\sigma_{\mu\nu} = 0$, as imposed by homogeneity and isotropy.

3.1.4.3 Recession velocity and proper velocity

Since the physical distance r is related to the comoving distance x with the relation $r = a(t)x$, we deduce that

$$\dot{r} = Hr + a(t)\dot{x} \equiv v_{\text{rec}} + v_{\text{proper}}. \quad (3.20)$$

The recession velocity, v_{rec} , is a time-dependent quantity and represents the component of the velocity related to the Hubble flow. As for v_{proper} , it is the proper velocity of the object with respect to the cosmological reference frame privileged by the expansion.

3.1.4.4 Properties of light

In the geometrical optics limit, it was shown that the trajectory of a light ray, $x^\alpha(\lambda)$, is given by a null geodesic with wavevector $k^\alpha = dx^\alpha/d\lambda$ satisfying

$$k^\mu k_\mu = 0 \quad \text{and} \quad k^\mu \nabla_\mu k^\nu = 0. \quad (3.21)$$

Using the general decomposition (3.18), we obtain

$$k^\mu = E u^\mu + p^\mu,$$

where $E = -k^\mu u_\mu$ represents the energy of a photon measured by the comoving observer and p^μ its momentum ($p^\mu u_\mu = 0$). It follows that

$$-E^2 + a^2 \gamma_{ij} p^i p^j = 0,$$

which is simply the generalization of the well-known relation between the energy of an object, its momentum and its mass. Moreover, the projection along u^μ of the geodesic equation gives

$$E\dot{E} + \Gamma_{ij}^0 p^i p^j = E\dot{E} + Ha^2 \gamma_{ij} p^i p^j = 0,$$

so that $\dot{E}/E = -H$ and hence $E = k_0/a$, k_0 being a constant. The energy, and hence the frequency $\nu \propto E$, of any photon varies as the inverse of the scale factor. Its wavelength varies as the scale factor. In an expanding space, wavelengths are thus redshifted achromatically.

The redshift defined by the general equation (1.122) is therefore given by

$$1+z = \frac{(u^\mu k_\mu)_{\text{emission}}}{(u^\mu k_\mu)_{\text{observation}}} = \frac{a(t_{\text{observation}})}{a(t_{\text{emission}})}. \quad (3.22)$$

The redshift can be measured from the comparison of an observed spectrum to a laboratory spectrum, from which one can deduce how much the Universe has expanded since it was emitted. We stress at this point that most observations give an angular position on the celestial sphere and a redshift. The radial distance cannot be measured directly but is obtained from the redshift and requires the knowledge of the expansion law of the Universe, that is $a(t)$.

3.1.4.5 Behaviour of a massive test particle

Consider now the trajectory of a massive test particle with 4-velocity v^μ . It satisfies

$$v^\mu v_\mu = -1, \quad v^\mu \nabla_\mu v^\nu = 0. \quad (3.23)$$

v^μ can be split as $v^\mu = \alpha u^\mu + p^\mu$, where u^μ is the 4-velocity of the fundamental observers and p^μ satisfies $u^\mu p_\mu = 0$. It follows from $u_\mu u^\mu = v_\mu v^\mu = -1$ that $p_\mu p^\mu = p^2 = -1 + \alpha^2$. The geodesic equation for v^μ then leads to $-\alpha \dot{\alpha} = (\Theta/3) \hat{\gamma}_{\mu\nu} p^\mu p^\nu = H p^2$ so that $\alpha \dot{\alpha} + H(-1 + \alpha^2) = 0$ from which we obtain that $-1 + \alpha^2 \propto a^{-2}$.

We conclude that $p^2 \propto a^{-2}$ so that $\alpha \rightarrow 1$ and $v^\mu \rightarrow u^\mu$. The proper velocity of any massive test particle is damped as a^{-1} and its worldline tends to adjust to the Hubble flow, that is to the worldlines of the fundamental observers.

3.1.5 Space-time dynamics

Up to now we have described the implications of the Robertson–Walker geometry. To investigate the dynamics of these space-times, we must obtain the equations of motion deriving from the Einstein equation. When these equations are fulfilled, we refer to these solutions as Friedmann–Lemaître space-times.

3.1.5.1 Energy-momentum tensor

The energy-momentum tensor, $T_{\mu\nu}$, is a symmetric tensor of rank 2. The possible forms this tensor can take are reduced by the space-time symmetries. Indeed, the most general form compatible with symmetries is

$$T_{\mu\nu} = Au_\mu u_\nu + B\hat{\gamma}_{\mu\nu}.$$

B represents the only spatial component (for the fundamental observers); it is therefore identified with the pressure that is thus seen to reduce to a unique number for a homogeneous and isotropic space. $A = T_{\mu\nu}u^\mu u^\nu \equiv \rho$ is the energy density measured by the fundamental observers. In conclusion, the most general form of the energy-momentum tensor is that of a perfect fluid

$$T_{\mu\nu} = \rho u_\mu u_\nu + P\hat{\gamma}_{\mu\nu}. \quad (3.24)$$

The energy density, ρ , and the pressure, P , only depend on time. This form is completely fixed by the cosmological principle and is not an independent choice. Thus, the only remaining freedom is the choice of the equation of state, that is the relation between the pressure and the energy density.

3.1.5.2 Friedmann and conservation equations

Using the expressions for the Einstein tensor and the form (3.24) of the energy-momentum tensor, we can easily deduce that the two Einstein equations reduce to two independent equations

$$H^2 = \frac{\kappa}{3}\rho - \frac{K}{a^2} + \frac{\Lambda}{3}, \quad (3.25)$$

$$\frac{\ddot{a}}{a} = -\frac{\kappa}{6}(\rho + 3P) + \frac{\Lambda}{3}, \quad (3.26)$$

recalling [see (1.130)] that $\kappa \equiv 8\pi G_N$.

As for the matter conservation equation ($\nabla_\mu T^{\mu\nu} = 0$), it reduces to a single equation

$$\dot{\rho} + 3H(\rho + P) = 0. \quad (3.27)$$

We can convince ourselves that the three equations (3.25), (3.26) and (3.27) are not independent [indeed, by differentiating (3.25) with respect to time and expressing \ddot{a}/a using (3.26) and (3.25) to eliminate H^2 , we find (3.27)]. This is a consequence of the Bianchi identities.

3.1.5.3 Covariant approach

In the covariant approach, a Robertson–Walker space is defined by the conditions $\bar{\nabla}_\mu f = 0$ for any scalar function f , which implies that $u^\mu = 0$, and $\sigma_{\mu\nu} = \omega^\mu = 0$, so that $\nabla_\mu u_\nu = \Theta\hat{\gamma}_{\mu\nu}/3$. Space being isotropic, we can set $S = a$, and thus $H = \Theta/3$ and the Raychaudhuri equation (1.73) gives

$$3\frac{\ddot{a}}{a} = -\frac{1}{2}R_{\mu\nu}u^\mu u^\nu = -\frac{\kappa}{2}(\rho + 3P) + \Lambda,$$

and the generalized Friedmann equation (1.78) gives⁴

$$(3) \hat{R} = \frac{6K}{a^2} = 2\kappa\rho + 2\Lambda - \frac{2}{3}\Theta^2 = 2\kappa\rho + 2\Lambda - 6H^2.$$

We recover the Friedmann equations, but the regime of validity of the Raychaudhuri and generalized Friedmann equations is larger since they also apply to non-isotropic Universes, such as Bianchi Universes (see below).

3.1.5.4 Equations in conformal time

It will sometimes be convenient to work in terms of the conformal time. Remembering that $dt = ad\eta$, we find that the derivative of any quantity X with respect to η , X' , is related to that with respect to t by $X' = a\dot{X}$. It will be convenient to remember that

$$X' = a\dot{X}, \quad X'' - \mathcal{H}X' = a^2\ddot{X}.$$

The comoving Hubble constant $\mathcal{H} = a'/a$ has been introduced, which satisfies in particular

$$\mathcal{H} = aH, \quad a\ddot{a} = \frac{a''}{a} - \mathcal{H}^2, \quad a^2\dot{H} = \mathcal{H}' - \mathcal{H}^2.$$

We deduce that the Friedmann and conservation equations [(3.25) to (3.27)] take the form

$$\boxed{\mathcal{H}^2 = \frac{\kappa}{3}\rho a^2 - K + \frac{\Lambda}{3}a^2,} \quad (3.28)$$

$$\boxed{\mathcal{H}' = -\frac{\kappa}{6}a^2(\rho + 3P) + \frac{\Lambda}{3}a^2,} \quad (3.29)$$

$$\boxed{\rho' + 3\mathcal{H}(\rho + P) = 0.} \quad (3.30)$$

Notice also that if the scale factor is a power law in cosmic time, then

$$a(t) \propto t^n \iff a(\eta) \propto \eta^{\frac{n}{1-n}}, \quad (3.31)$$

as long as $n \neq 1$.

3.1.5.5 Equation of state

We hence have two independent equations for three unknowns (the scale factor, the energy density and the pressure). We therefore need some additional information to solve this system. For that, the most convenient way is to give an equation of state for the matter in the form

$$P = w\rho. \quad (3.32)$$

For instance, pressureless matter will be described by $w = 0$ whereas for radiation, we will have $w = 1/3$. For the cosmological constant, Λ , which corresponds to a constant

⁴Note that here $(3)\hat{R}$ refers to the Gaussian curvature of the hypersurfaces Σ_t with metric $\hat{\gamma}_{\mu\nu}$. This explains why $(3)\hat{R} = 6K/a^2$.

energy density, (3.27) implies $P = -\rho$ and thus $w = -1$. The resolution of (3.27) implies that for any constant w ,

$$\rho \propto a^{-3(1+w)}. \quad (3.33)$$

Comparing this general result with equations (3.25) and (3.26), we notice that the curvature term can be assimilated with a fluid with equation of state $w = -1/3$.

By differentiating w with respect to time and expressing ρ' with the conservation equation (3.30) we get

$$w' = -3H(1+w)(c_s^2 - w), \quad (3.34)$$

where c_s is the speed of sound, defined by

$$c_s^2 = \frac{P'}{\rho'}. \quad (3.35)$$

For a barotropic fluid $w = c_s^2$, so that w is a constant.

3.1.5.6 Reduced form

It is useful to rewrite the Friedmann equations in a dimensionless form in terms of reduced quantities. For that, we introduce the energy density parameters

$$\Omega = \frac{\kappa\rho}{3H^2}, \quad \Omega_\Lambda = \frac{\Lambda}{3H^2}, \quad \Omega_K = -\frac{K}{H^2a^2}, \quad (3.36)$$

respectively, for the matter, the cosmological constant and the curvature. The matter term can be decomposed as a sum of components with different equations of state as $\Omega = \sum_x \Omega_x$ with

$$\Omega_x = \frac{\kappa\rho_x}{3H^2}, \quad (3.37)$$

and we will denote by Ω_{x0} its value today. The first Friedmann equation (3.25) then takes the form of a constraint

$$\sum_x \Omega_x + \Omega_\Lambda + \Omega_K = 1. \quad (3.38)$$

Using the solutions of the conservation equation, it follows that

$$\Omega_x = \Omega_{x0} \left(\frac{a}{a_0} \right)^{-3(1+w_{x0})} \left(\frac{H_0}{H} \right)^2,$$

for a constant equation of state⁵ $w_x = w_{x0}$, so that $E(a) \equiv H/H_0$ can be expressed as

⁵For a general time-dependent equation of state, it can be checked that $\Omega_x = \Omega_{x0} (H_0/H)^2 \exp \left[-3 \int_{a_0}^a (1+w_x) d \ln a \right]$.

$$E^2(a) \equiv \left(\frac{H}{H_0}\right)^2 = \sum_x \Omega_{x0} \left(\frac{a}{a_0}\right)^{-3(1+w_x)} + \Omega_{K0} \left(\frac{a}{a_0}\right)^{-2} + \Omega_{\Lambda0}. \quad (3.39)$$

Quantities with an index 0 are evaluated today. Thus, t_0 will be the dynamical age of the Universe and $a_0 = a(t_0)$. Notice that E only depends on $x = a/a_0 = 1/(1+z)$. Depending on the case, it will be denoted by $E(a)$, $E(x)$ or $E(z)$.

For the matter, the following different components can be distinguished

| | |
|------------------------------|--|
| Total matter | $\overbrace{\hspace{10em}}^{\Omega}$ |
| Different equations of state | $\overbrace{\hspace{3em}}^{\Omega_m} \quad \overbrace{\hspace{3em}}^{\Omega_r} \quad \overbrace{\hspace{3em}}^{\Omega_\Lambda}$ |
| Different species | $\overbrace{\hspace{2em}}^{\Omega_b} \quad \overbrace{\hspace{2em}}^{\Omega_c} \quad \overbrace{\hspace{2em}}^{\Omega_\gamma} \quad \overbrace{\hspace{2em}}^{\Omega_\nu}$ |

where the total matter component is decomposed into matter ($w = 0$) and radiation ($w = 1/3$) and then, respectively, into baryons (b), cold dark matter (c), photons (γ) and neutrinos (ν). Of course, other species can be added, whether as a species with a new equation of state or as a new component.

3.1.5.7 Units and conventions

Two conventions are often used for the choice of dimensions. The decomposition of any physical length as the product of the scale factor and a comoving length, $\ell^{(\text{phys})} = a(t)\ell^{(\text{com})}$, can indeed lead to an ambiguity since we can arbitrarily choose either the scale factor or the comoving length to have dimension of length. We will use both conventions according to the situation at hand.

- The scale factor can be chosen to have dimension of length, and χ to be dimensionless. This implies that K can be normalized to $K = 0, \pm 1$. The value of a_0 can then be determined by evaluating the Friedmann equations today.

$$[a] = L \text{ and } [\chi] = 1: \chi \equiv \bar{\chi}, \quad K = 0, \pm 1, \quad a_0 = \frac{1}{H_0} \frac{1}{\sqrt{|1 - \Omega_0|}}.$$

For a Universe such that $\Omega_0 = 1$ (i.e. $K = 0$), a_0 is arbitrary, which reflects the invariance of the Euclidean space under dilatation.

- The scale factor can be chosen to be dimensionless. χ will then have dimension of length, and thus K the dimension of inverse length squared. It is then often convenient to set $a_0 = 1$.

$$[a] = 1 \text{ and } [\chi] = L: K = (0, \pm 1)/R_c^2, \quad R_c = \frac{1}{a_0 H_0} \frac{1}{\sqrt{|1 - \Omega_0|}}.$$

We recover that for $\Omega_0 = 1$ (i.e. $K = 0$), $R_c = \infty$. A Euclidean space has no characteristic curvature scale.

3.2 Dynamics of Friedmann–Lemaître space-times

3.2.1 Some solutions

Using the Friedmann equation (3.25), the evolution law for the scale factor can be determined as a function of time for some simple cases.

3.2.1.1 $K = 0$ and $w \neq -1$

As seen earlier, if $w \neq -1$, the conservation equation implies that $\rho \propto a^{-3(1+w)}$, so that the Friedmann equation (3.25) implies $H^2 \propto a^{-3(1+w)}$. It follows that $\dot{a} \propto a^{-(1+3w)/2}$. By integrating this relation, we obtain,

$$a(t) \propto t^{\frac{2}{3(1+w)}}. \quad (3.40)$$

From this expression, the conformal time can be computed by integrating $d\eta = dt/a(t)$, giving $\eta \propto t^{\frac{1+3w}{3(1+w)}}$ if $w \neq -1/3$. We deduce that

$$a(\eta) \propto \eta^{\frac{2}{1+3w}} \quad \text{if } w \neq -\frac{1}{3}, \quad (3.41)$$

and $a(\eta) \propto \exp \eta$ otherwise.

3.2.1.2 $K = 0$ and $w = -1$

If the matter density is dominated by a cosmological constant, we have $H^2 \propto \rho \propto a^0$, so that $\dot{a} \propto a$. By integrating this relation, we get

$$a(t) \propto e^{Ht}. \quad (3.42)$$

This space has an accelerated expansion and is called de Sitter space. In terms of the conformal time, the scale factor is of the form

$$a(\eta) \propto -\frac{1}{H\eta}, \quad \eta < 0. \quad (3.43)$$

This space can also be written in such a way that it has spherical spatial sections (see Chapter 8).

$$ds^2 = -dt^2 + \frac{\cosh^2(Ht)}{H^2} (d\chi^2 + \sin^2 \chi d\Omega^2). \quad (3.44)$$

3.2.1.3 $K = \pm 1$ and $w \neq -1/3$

It is more convenient to work in conformal time to obtain $a(t)$ in the parametric form $\{a(\eta), t(\eta)\}$. Using $\rho = \rho_0(a/a_0)^{-3(1+w)}$ and (3.25), we get

$$\mathcal{H}^2 = \frac{\Omega_0}{|1 - \Omega_0|} \left(\frac{a}{a_0} \right)^{-(1+3w)} - K,$$

in units such that $|K| = 1$. This equation can be integrated to give

$$\left(\frac{a}{a_0}\right) = \left(\frac{\Omega_0}{|1 - \Omega_0|}\right)^{1/2\alpha} \begin{cases} (\sinh \alpha\eta)^{1/\alpha} & \text{if } K = -1, \\ (\sin \alpha\eta)^{1/\alpha} & \text{if } K = +1, \end{cases} \quad (3.45)$$

with $2\alpha = 1 + 3w$. This solution is valid for all constant $w \neq -1/3$. The case $w = -1/3$ corresponds to a Universe with no matter since matter can then be absorbed into the curvature term.

3.2.1.4 $K = 0, w = 0$ and $\Lambda \neq 0$

Let us mention the solution for a spatially Euclidean model dominated by a pressureless fluid and a cosmological constant [11]

$$a(t) = \left(\frac{1}{\Omega_{\Lambda 0}} - 1\right)^{1/3} \sinh^{2/3}\left(\frac{3\alpha t}{2}\right), \quad (3.46)$$

where we $\alpha = H_0\sqrt{\Omega_{\Lambda 0}}$. This solution is relevant for describing the late time evolution of a flat Λ CDM model.

3.2.1.5 Other solutions

We present in Fig. 3.3 the profile of the scale-factor evolution in a large number of situations, including for different kinds of matter, curvature and value of the cosmological constant.

We may notice the presence of periods of slower evolution (a ‘hesitating’ phase) of various lengths depending on the values of the cosmological constant (see, e.g., cases F, L or K). During these phases, the Universe gets closer to an Einstein static Universe (i.e. $\dot{a} = 0$) for which

$$a^2 = 1/\Lambda_E, \quad \kappa\rho = 2\Lambda_E.$$

Thus, this is a spherical Universe for which the gravitational attraction is just compensated by the repulsive effect of the cosmological constant. The behaviour of these solutions depends mainly on the relative values of the cosmological constant and Λ_E .

Another interesting solution is the one obtained for an empty Universe with hyperbolic spatial sections ($K = -1$). This space, called the Milne space, has the following metric

$$ds^2 = -dt^2 + t^2(d\chi^2 + \sinh^2\chi d\Omega^2). \quad (3.47)$$

It can be shown that this metric corresponds to a Minkowski metric by an appropriate change of coordinates ($T = t \cosh \chi, R = t \sinh \chi$ leads to the Minkowski metric in spherical coordinates), but this space only covers a quarter of the Minkowski space-time.

3.2.2 Dynamical evolution

In order to study the general dynamics of the solutions of the Friedmann equations, it is interesting to rewrite the system (3.25) and (3.26) in the form of a dynamical system for the quantities Ω , Ω_Λ and Ω_K (see Refs. [12–14]).

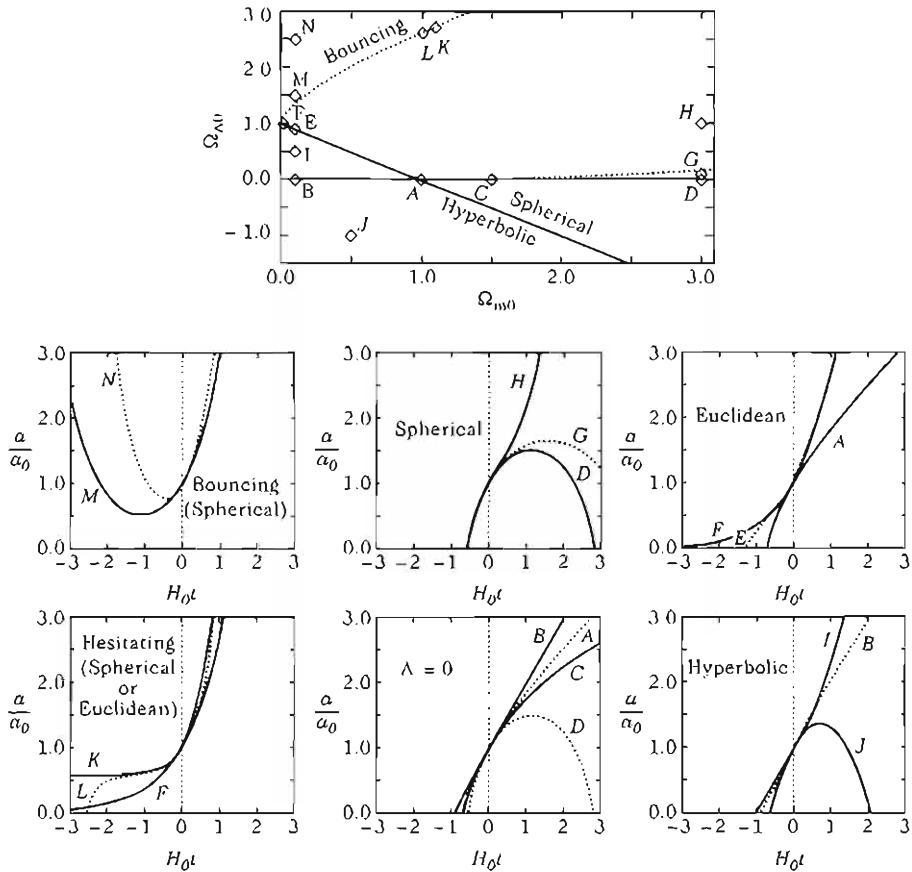


Fig. 3.3 Depending on the values of the cosmological parameters (upper panel), the scale factor can have very different evolutions.

3.2.2.1 Dynamical system

Using $\dot{H} = \ddot{a}/a - H^2$ and expressing \ddot{a}/a with (3.26) and H^2 with (3.25) leads to

$$\frac{\dot{H}}{H^2} = -(1 + q). \quad (3.48)$$

q is the deceleration parameter defined by

$$2q \equiv (3\gamma - 2)(1 - \Omega_K) - 3\gamma\Omega_\Lambda, \quad (3.49)$$

with $\gamma \equiv w + 1$.

The system of Friedmann equations can thus be rewritten using the new time variable $p \equiv \ln(a/a_0)$. The derivative, X' , with respect to p of any quantity X is then given by $X' = X/H$.

The evolution equation for the Hubble parameter (3.48) then becomes

$$H' = -(1+q)H. \quad (3.50)$$

If we differentiate Ω , Ω_Λ and Ω_K with respect to p and use (3.50) to express H' , noticing that $a' = a$ and that (3.30) can be used to obtain $\rho' = -3\gamma\rho$, we then obtain the system

$$\Omega' = (2q + 2 - 3\gamma)\Omega, \quad (3.51)$$

$$\Omega'_\Lambda = 2(1+q)\Omega_\Lambda, \quad (3.52)$$

$$\Omega'_K = 2q\Omega_K. \quad (3.53)$$

It is useless to keep all of these equations since (i) H does not enter in (3.51)–(3.53) and (ii) Ω can be deduced algebraically using (3.38).

The dynamics of Friedmann–Lemaître space-times can thus be studied by considering the dynamical system

$$\begin{cases} \Omega'_\Lambda = 2(1+q)\Omega_\Lambda, \\ \Omega'_K = 2q\Omega_K, \end{cases} \quad (3.54)$$

where q is a function of Ω_Λ and Ω_K given by (3.49).

This system associates a unique tangent vector to each point so that only one integral curve passes through each point of the phase space where the tangent vector is defined. If two trajectories crossed each other, the tangent vector at this point would not be unique. Points where the tangent vector is not defined are called *fixed points*.

3.2.2.2 Determination of the fixed points

The fixed points are defined by the conditions $\Omega'_\Lambda = 0$ and $\Omega'_K = 0$ and are thus solutions of

$$(1+q)\Omega_\Lambda = 0, \quad q\Omega_K = 0. \quad (3.55)$$

They represent the equilibrium positions (stable or not). This equation has three solutions

$$(\Omega_K, \Omega_\Lambda) \in \{(0, 0), (0, 1), (1, 0)\}, \quad (3.56)$$

and each of these solutions represent a Universe with different characteristics.

- $(\Omega_K, \Omega_\Lambda) = (0, 0)$: the *Einstein-de Sitter space* (EdS). This is a space-time with no curvature (Euclidean spatial sections) and no cosmological constant.
- $(\Omega_K, \Omega_\Lambda) = (0, 1)$: the *de Sitter space* (dS). This is a space with no matter but with a cosmological constant and the spatial sections have a positive curvature. This Universe is in eternal exponential expansion.
- $(\Omega_K, \Omega_\Lambda) = (1, 0)$: the *Milne space* (M). This is an empty space with no cosmological constant but with hyperbolic spatial sections ($K < 0$).

3.2.2.3 Stability of the fixed points

To determine if these fixed points are attractors (A), saddle points (S), or repellers (R), we should study the evolution of a small perturbation around each equilibrium position. For this we set

$$\Omega_K \equiv \bar{\Omega}_K + \omega_K, \quad (3.57)$$

$$\Omega_\Lambda \equiv \bar{\Omega}_\Lambda + \omega_\Lambda, \quad (3.58)$$

with $(\bar{\Omega}_\Lambda, \bar{\Omega}_K)$ the coordinates of a fixed point and $(\omega_\Lambda, \omega_K)$ a small deviation.

Rewriting the system (3.54) in the form

$$\begin{pmatrix} \Omega_K \\ \Omega_\Lambda \end{pmatrix}' = \begin{bmatrix} F_K(\Omega_\Lambda, \Omega_K) \\ F_\Lambda(\Omega_\Lambda, \Omega_K) \end{bmatrix}, \quad (3.59)$$

where F_K and F_Λ are two functions determined by the system (3.54) that vanish at the fixed point, it can be expanded to linear order as

$$\begin{pmatrix} \omega_K \\ \omega_\Lambda \end{pmatrix}' = P_{(\bar{\Omega}_\Lambda, \bar{\Omega}_K)} \begin{pmatrix} \omega_K \\ \omega_\Lambda \end{pmatrix} \quad \text{with} \quad P_{(\bar{\Omega}_\Lambda, \bar{\Omega}_K)} \equiv \begin{pmatrix} \frac{\partial F_K}{\partial \Omega_K} & \frac{\partial F_K}{\partial \Omega_\Lambda} \\ \frac{\partial F_\Lambda}{\partial \Omega_K} & \frac{\partial F_\Lambda}{\partial \Omega_\Lambda} \end{pmatrix}_{(\bar{\Omega}_\Lambda, \bar{\Omega}_K)}. \quad (3.60)$$

The stability of a fixed point depends on the sign of the two eigenvalues $(\lambda_{1,2})$ of the matrix $P_{(\bar{\Omega}_\Lambda, \bar{\Omega}_K)}$. If both eigenvalues are positive (resp., negative), the fixed point is unstable (resp., stable). If the two eigenvalues have different sign, it is a saddle point. The eigenvector $u_{\lambda_{1,2}}$ then gives the stable and unstable directions.

– EdS: we have

$$P_{\text{EdS}} = \begin{pmatrix} 3\gamma - 2 & 0 \\ 0 & 3\gamma \end{pmatrix},$$

with eigenvalues $3\gamma - 2$ and 3γ . EdS is thus an attractor for $\gamma \in]-\infty, 0[$, a saddle point for $\gamma \in]0, 2/3[$ and a repeller for $\gamma \in]2/3, +\infty[$. The associated eigendirections are

$$u_{(3\gamma)} = (0, 1), \quad u_{(3\gamma-2)} = (1, 0).$$

– dS: we have

$$P_{\text{dS}} = \begin{pmatrix} -2 & 0 \\ 2 - 3\gamma & -3\gamma \end{pmatrix},$$

with eigenvalues -2 and -3γ . If $\gamma \in]-\infty, 0[$ then dS is a saddle point and for $\gamma \in]0, +\infty[$, it is an attractor. The two eigendirections are given by

$$u_{(-3\gamma)} = (0, 1), \quad u_{(-2)} = (1, -1).$$

– M: we have

$$P_M = \begin{pmatrix} 2 - 3\gamma & -3\gamma \\ 0 & 2 \end{pmatrix},$$

with eigenvalues 2 and $2 - 3\gamma$. If $\gamma \in]-\infty, 2/3[$ then M is a repeller and if $\gamma \in]2/3, +\infty[$, it is a saddle point. The two associated eigenvectors are

$$u_{(2-3\gamma)} = (1, 0), \quad u_{(2)} = (1, -1).$$

These results are summarised in Table 3.1 and in Figs. 3.4 to 3.6.

Table 3.1 Stability of the fixed points as a function of the value of the equation of state, $\gamma = w + 1$, (A : attractor, R : repeller, and S : saddle point).

| γ | $]-\infty, 0[$ | 0 | $]0, 2/3[$ | $2/3$ | $]2/3, +\infty[$ |
|----------|----------------|------|------------|-------|------------------|
| EdS | A | N.A. | S | N.A. | R |
| dS | S | A | A | A | A |
| M | S | R | R | R | S |

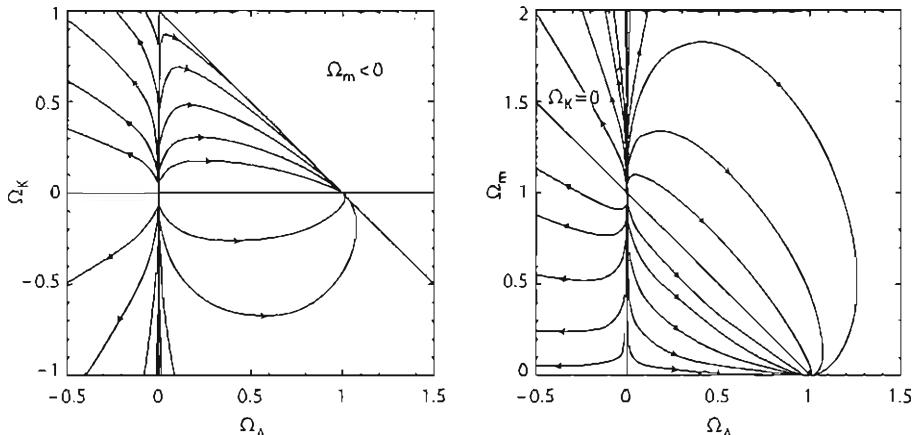


Fig. 3.4 Dynamical evolution in the plane $(\Omega_K, \Omega_\Lambda)$ (left) and $(\Omega_m, \Omega_\Lambda)$ (right) for $\gamma = 1$ ($w = 0$). From Ref. [14].

3.2.3 Expansion and contraction

When the total energy density vanishes, the expansion rate of the Universe can change sign. For a Universe with non-flat spatial sections and with matter, radiation and a cosmological constant, there are four possibilities (see Fig. 3.3).

- The expansion lasts eternally starting from the Big Bang. This is the case if $K = 0$ or -1 for a Universe without cosmological constant.
- The expansion is followed by a contracting phase leading to a big crunch. This is the case for a Universe with $K = +1$ with no cosmological constant.
- The expansion lasts eternally but was preceded by a contracting phase; there must therefore have been a *bounce*. This situation can occur when the Universe has spherical spatial sections ($K = +1$).
- The Universe undergoes a series of contraction and expansion phases; this is an oscillating Universe.

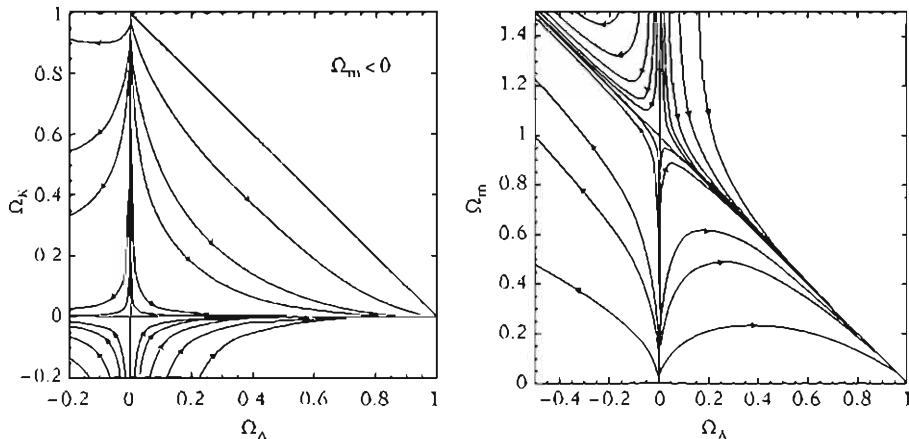


Fig. 3.5 Dynamical evolution in the plane $(\Omega_K, \Omega_\Lambda)$ (left) and $(\Omega_m, \Omega_\Lambda)$ (right) for $\gamma = 1/3$. From Ref. [14].

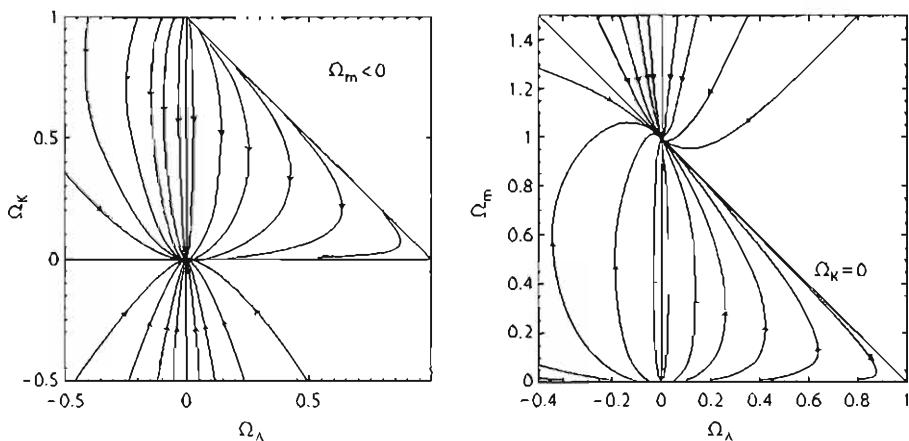


Fig. 3.6 Dynamical evolution in the plane $(\Omega_K, \Omega_\Lambda)$ (left) and $(\Omega_m, \Omega_\Lambda)$ (right) for $\gamma = -1$. From Ref. [14].

The separation between a Universe with infinite expansion and a Universe with a big crunch is obtained when

$$\Omega_{m0} \leq 1, \quad \Omega_{\Lambda0} = 0,$$

or

$$\Omega_{m0} \geq 1, \quad \Omega_{\Lambda0} = 1 - \Omega_{m0} + \frac{2}{3} \cos [\arctan (\Xi_{m0}^{-1}) - \pi], \quad (3.61)$$

with $\Xi_{m0} = \sqrt{|2\Omega_{m0} - 1|}$. The separation between a Universe with infinite expansion and a bouncing Universe is obtained when

$$\Omega_{m0} = 0, \quad \Omega_{\Lambda0} \leq 0,$$

or

$$0 \leq \Omega_{m0} \leq \frac{1}{2}, \quad \Omega_{\Lambda0} = 1 - \Omega_{m0} + \frac{3}{2(\Omega_{m0})^{1/3}} \left[(1 - \Omega_{m0} + \Xi_{m0})^{1/3} + (1 - \Omega_{m0} - \Xi_{m0})^{1/3} \right],$$

or

$$\Omega_{m0} \geq \frac{1}{2}, \quad \Omega_{\Lambda0} = 1 - \Omega_{m0} + \frac{2}{3} \cos [\arctan (\Xi_{m0}^{-1}) + \pi]. \quad (3.62)$$

3.3 Time and distances

The characteristic distance and time scales of any Friedmann–Lemaître Universe are fixed by the value of the Hubble constant. We define the Hubble time and distance by

$$t_H = \frac{1}{H}, \quad D_H = \frac{c}{H}. \quad (3.63)$$

The sphere with radius equal to the Hubble radius is called the *Hubble sphere*. The order of magnitude of these quantities is obtained by expressing the current value of the Hubble parameter in the following way

$$H_0 = 100 h \text{ km} \cdot \text{s}^{-1} \cdot \text{Mpc}^{-1}, \quad (3.64)$$

with h typically of the order of 0.7. We will come back later to the measurement of all the parameters introduced in this chapter. We obtain that

$$D_{H_0} = 9.26 h^{-1} \times 10^{25} \text{ m} \sim 3000 h^{-1} \text{ Mpc}, \quad t_{H_0} = 9.78 h^{-1} \times 10^9 \text{ years}. \quad (3.65)$$

All the quantities that will be introduced in this section (distance and time) will be expressed in terms of $E(z)$ defined in (3.39). As can be seen in Fig. 3.7, this function has some degeneracies that change with the redshift. By combining observations at different redshifts, one can lift these degeneracies and thus use them jointly to better constrain the cosmological parameters.

3.3.1 Age of the Universe and look-back time

From the definition of the Hubble parameter, it follows that $dt = da/aH$. Thus, the expression (3.39) for the Hubble constant gives

$$dt = t_{H_0} \frac{da}{aE(a)}.$$

The age of the Universe is thus obtained by integrating this equality between $a = 0$ and $a = a_0$, or equivalently between $z = 0$ and $z = \infty$,

$$t_0 = t_{H_0} \int_0^1 \frac{dx}{xE(x)} = t_{H_0} \int_0^\infty \frac{dz}{(1+z)E(z)}. \quad (3.66)$$

This function is represented for various sets of cosmological parameters in Fig. 3.8.

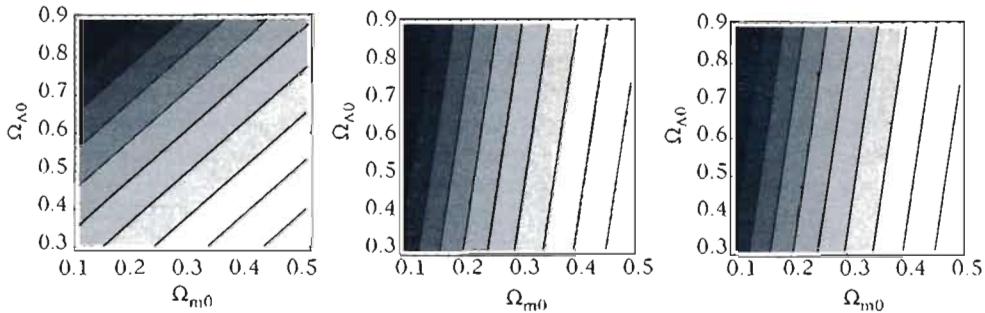


Fig. 3.7 Isocontours of the function $E(z)$ in terms of the cosmological parameter $(\Omega_{m0}, \Omega_{k0})$ for a Λ CDM model. The degeneracies at $z = 1$ (left), $z = 3$ (middle) and $z = 10$ (right) are not the same. The code of grey shade is arbitrary.

The analogous expression in conformal time is

$$\eta_0 = t_{H_0} \int_0^1 \frac{dx}{x^2 E(x)}. \quad (3.67)$$

In a similar way, the age of the Universe at the time when a photon with redshift z_* was emitted is defined by

$$t(z_*) = t_{H_0} \int_{z_*}^{\infty} \frac{dz}{(1+z)E(z)}. \quad (3.68)$$

The look-back time is the difference between the age of the Universe and its age when the photon was emitted by the source. The previous relations lead to

$$\Delta t(z_*) = t_0 - t(z_*) = t_{H_0} \int_0^{z_*} \frac{dz}{(1+z)E(z)}. \quad (3.69)$$

3.3.2 Comoving radial distance

The comoving radial distance χ of an object of redshift z that is observed by an observer located at $\chi = 0$, is obtained by integrating along a radial null geodesic. The form of the metric (3.4) gives $d\chi = dt/a$, so that

$$a_0 d\chi = \frac{dx}{x^2 H(x)} = -\frac{dz}{H(z)}.$$

It follows that

$$a_0 \chi(z_*) = D_{H_0} \int_0^{z_*} \frac{dz}{E(z)}. \quad (3.70)$$

The dependance on a_0 simply reflects the arbitrary choice of the system of units. This function is represented for various sets of cosmological parameters in Fig. 3.9.

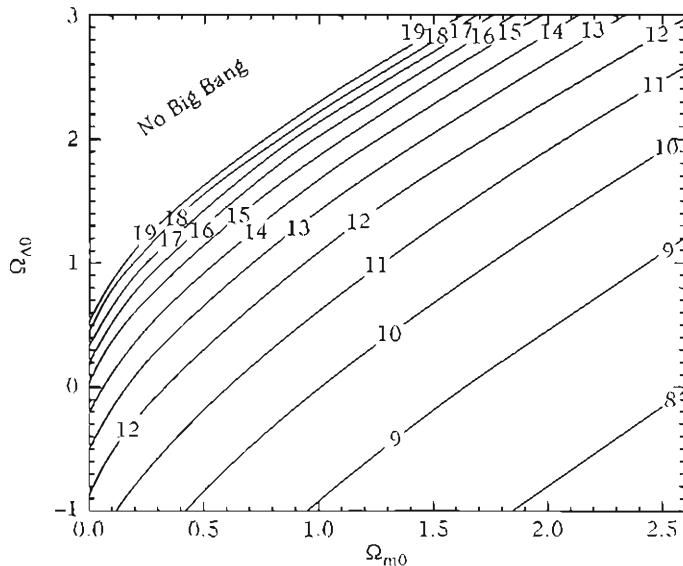


Fig. 3.8 Isocontours of the function t_0 as a function of the cosmological parameters. Each curve indicates the value of t_0 in units of 10^9 years. The grey zone represents the zone of parameters for which there is no Big Bang.

For a flat Universe with no cosmological constant, we find, for instance,

$$a_0 \chi(z) = 2D_{H_0} \left(1 - \frac{1}{\sqrt{1+z}} \right). \quad (3.71)$$

For a Universe with a non-vanishing curvature, we can give the expression for the radial distance in units of the curvature radius

$$\bar{x}(z_s) = \sqrt{|1-\Omega_0|} \int_0^{z_s} \frac{dz}{E(z)}. \quad (3.72)$$

3.3.3 Angular distances

3.3.3.1 Comoving angular diameter

The comoving angular diameter relates the comoving transverse size of an object and the solid angle under which it is observed. It is defined by the relation

$$dS_{\text{source}}^{\text{com}} = R_{\text{ang}}^2(\chi) d\Omega_{\text{obs}}^2,$$

between the solid angle and the apparent surface of an object. Remembering that a comoving sphere centred at $\chi = 0$ and with comoving radius χ has a comoving surface $S^{(\text{com})} = 4\pi f_K^2(\chi)$, we deduce that

$$R_{\text{ang}}(z) = f_K[\chi(z)]. \quad (3.73)$$

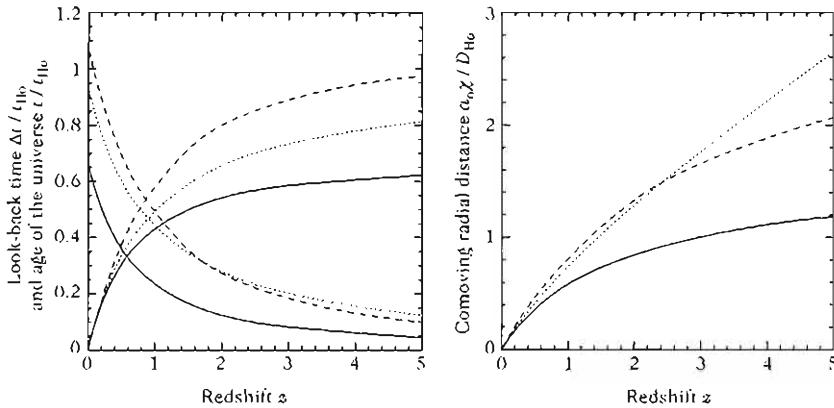


Fig. 3.9 (left): Age and look-back time in units of the Hubble time, $t(z)/t_{H_0}$ and $\Delta t(z)/t_{H_0}$ as a function of z for three cosmological models defined by $(\Omega_{m0}, \Omega_{\Lambda0}) = (1, 0)$, $(0.05, 0)$ and $(0.2, 0.8)$, respectively, as plain, dotted and dashed lines. (right): Comoving radial distance in units of the Hubble length, $a_0\chi(z)/D_{H_0}$, as a function of z for the same cosmological models.

Notice that this function is not necessarily increasing with z . Indeed, if $K = +1$, then two objects with the same comoving size, located at χ and $\pi - \chi$ will have the same comoving angular diameter.

3.3.3.2 Angular distance

The angular distance is the notion that generalizes the concept of parallax. It is defined, for a geodesic bundle converging at the point where the observer is, as the ratio between the transverse physical size of an object and the solid angle under which it is observed. The physical area of the source is related to the comoving one by $dS^{\text{phys}} = a^2 dS^{\text{com}}$. With this definition

$$D_A^2 = \frac{dS_{\text{source}}^{\text{phys}}}{d\Omega_{\text{obs}}^2} = a^2 \frac{dS_{\text{source}}^{\text{com}}}{d\Omega_{\text{obs}}^2}.$$

This quantity will also be denoted by r_o . Using the definition of the comoving angular diameter, introduced in the previous section, we obtain

$$r_o(z) = D_A(z) = a_0 \frac{f_K |\chi(z)|}{1+z} = \frac{a_0 R_{\text{ang}}(z)}{1+z}.$$

(3.74)

3.3.3.3 Reciprocity theorem

In the same way, we can define the angular distance from the source. It is defined using a geodesic bundle diverging from the source as the ratio between the physical transverse size of the observer and the solid angle under which it would be observed from the source,

$$dS_{\text{obs}}^{\text{phys}} = r_s^2 d\Omega_{\text{source}}^2.$$

It can be shown that if the trajectories of the photons are null geodesics and if the equation for the geodesic deviation is valid, then there exists a relation between r_s and r_o

$$r_s = r_o(1 + z), \quad (3.75)$$

called the *reciprocity theorem* (see Ref. [15] for a demonstration). This relation cannot be directly checked since r_s cannot be observed directly.

3.3.4 Luminosity distance

3.3.4.1 Definition

The luminosity distance relates the luminosity of a source, located at a comoving radial distance χ from the observer, to the observed flux as

$$\phi_{\text{obs}} = \frac{L_{\text{source}}(\chi)}{4\pi D_L^2}. \quad (3.76)$$

To relate the luminosity of the source and the observed luminosity, notice that the luminosity of the source is given by the emitted energy per unit of emission time, $L_{\text{source}} = \Delta E_{\text{em}}/\Delta t_{\text{em}}$. Due to the expansion of the Universe, $\Delta E_{\text{obs}} = \Delta E_{\text{em}}a/a_0$ and $\Delta t_{\text{obs}} = \Delta t_{\text{em}}a_0/a$ (because $\Delta\eta_{\text{obs}} = \Delta\eta_{\text{em}}$ for two null geodesics) so that $L_{\text{obs}} = L_{\text{source}}(1+z)^{-2}$. The flux received by the observer at $\chi = 0$ is given by

$$\phi_{\text{obs}} = \frac{L_{\text{obs}}}{S^{(\text{phys})}} = \frac{L_{\text{source}}(1+z)^{-2}}{4\pi a_0^2 f_K^2(\chi)},$$

using $S^{(\text{phys})} = a_0^2 S^{(\text{com})}$. Identifying this with the definition (3.76), it follows that

$$D_L(z) = a_0(1+z)f_K[\chi(z)]. \quad (3.77)$$

The luminosity of the source is indeed unknown. However, by comparing some sources to a reference source for which the redshift z_* and the distance are known, we obtain that

$$\frac{D_L(z)}{D_L(z_*)} = \sqrt{\frac{L}{L_*} \frac{\phi_*}{\phi}}.$$

By extracting a family of astrophysical objects, called *standard candles*, with the same luminosity, the ratio of their fluxes can be used to measure the luminosity distance.

To finish, let us note that deriving (3.77) allows us to extract the Hubble rate as a function of redshift as

$$\frac{1}{H(z)} = \left[1 + \Omega_{K0} H_0^2 \left(\frac{D_L}{1+z} \right)^2 \right]^{-1/2} \frac{d}{dz} \left(\frac{D_L}{1+z} \right).$$

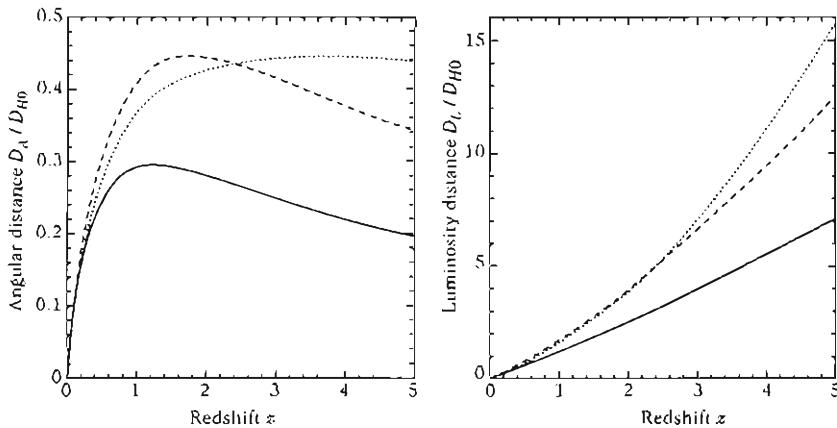


Fig. 3.10 (left): The angular distance $D_A(z)/D_{H_0}$ as a function of the redshift z for three cosmological models, defined by $(\Omega_{m0}, \Omega_{\Lambda0}) = (1, 0), (0.05, 0)$ and $(0.2, 0.8)$, respectively, as plain, dotted and dashed lines. (right): The luminosity distance $D_L(z)/D_{H_0}$ as a function of the redshift z for the three same cosmological models.

3.3.4.2 Distances duality relation

If the number of photons is conserved, which is usually the case if the Maxwell equations are valid, and if the absorption from the medium in which they propagate is small, then (for a proof, see Ref. [15])

$$D_L = r_s(1 + z),$$

so that (3.75) implies a duality relation between the angular and luminosity distances

$$D_L = (1 + z)^2 D_A. \quad (3.78)$$

Unlike (3.75), this relation can be tested experimentally [16].

3.3.4.3 Distance modulus

The luminosity distance is usually expressed in terms of the *distance modulus*, $m - M$, defined such that it vanishes if the source is located conventionally⁶ at 10 pc,

$$m - M = -2.5 \log [\phi(z)/\phi(10 \text{ pc})]. \quad (3.79)$$

Using (3.76) and noticing that at 10 pc, the luminosity distance is also 10 pc, we obtain $m - M = 5 \log[D_L(z)/10 \text{ pc}]$. Using the expression $10 \text{ pc} = D_{H_0}/3 \times 10^8 h$, it can be deduced that

$$m - M = 25 + 5 \log \left[3000 \frac{D_L}{D_{H_0}} \right] - 5 \log h. \quad (3.80)$$

⁶Note that the -2.5 factor is arbitrary and was chosen to match the magnitudes of stars initially defined by Hipparcos.

m is the apparent magnitude and the absolute magnitude M is defined by comparison to the Solar luminosity

$$M = -2.5 \log \left(\frac{L}{L_\odot} \right) + 4.76 \quad (3.81)$$

where $L_\odot \sim 3.8 \times 10^{26}$ W is the Sun luminosity and $M_\odot = 4.76$ its magnitude. With such a definition, the brighter an object is, the weaker is its magnitude.

3.3.4.4 Magnitudes

The luminosities considered previously were the luminosities integrated over the entire electromagnetic spectrum, defining the *bolometric magnitude*. It is difficult to measure these bolometric magnitudes and, in general, some filters are used, selecting a specific frequency band. The magnitude M_X in the band X is then defined by

$$M_X = -2.5 \log \left(\frac{L_X}{L_{\odot X}} \right) + M_\odot X. \quad (3.82)$$

An example of different frequency bands used in astronomy is given in the table below.

| Band | Central | Width (nm) | M_\odot |
|-------------------|-----------------|------------|-----------|
| | wavelength (nm) | | |
| U (ultraviolet) | 365 | 66 | 5.61 |
| B (blue) | 445 | 94 | 5.48 |
| V (visible) | 551 | 88 | 4.64 |
| R (red) | 658 | 138 | 4.42 |
| I (infra-red) | 806 | 149 | 4.08 |
| J | 1200 | 213 | 3.64 |
| K | 2190 | 390 | 3.28 |
| Bolometric | | ∞ | 4.76 |

The flux detected in the frequency band $\Delta\nu_X$ was emitted in the frequency band $(1+z)\Delta\nu_X$. The flux received per frequency unit is thus

$$\phi_X(\nu) = (1+z) \frac{L_\nu[\nu(1+z)]}{4\pi D_L^2}, \quad (3.83)$$

where L_ν is the source luminosity per frequency band. It follows that

$$\phi_X(\nu) = \frac{(1+z)}{4\pi D_L^2} \int_{\Delta\nu_X} L_\nu[\nu(1+z)] d\nu. \quad (3.84)$$

The correction induced by the deformation of the spectrum does not appear when the source is at 10 pc so that the distance modulus in the band X defined by $m_X - M_X = -2.5 \log[\phi_X/\phi_X(10 \text{ pc})]$, is given by

$$m_X - M_X = 5 \log \left(\frac{D_L}{10 \text{ pc}} \right) + K. \quad (3.85)$$

The K -correction represents the distortion of the spectrum due to the cosmological expansion

$$K = -2.5 \log(1+z) - 2.5 \log \left\{ \frac{\int_{\Delta\nu_X} L_\nu[\nu(1+z)]d\nu}{\int_{\Delta\nu_X} L_\nu[\nu]d\nu} \right\}. \quad (3.86)$$

Finally, note that other effects of astrophysical origin, such as the absorption by the interstellar or intergalactic medium, can affect the apparent magnitude. We shall discuss them in the next chapter.

3.3.5 Volume and number counts

The geometric properties of space-time also play a role in the determination of the distribution of a family of objects (for instance, galaxies) in a given redshift band. Let us consider such a family with proper number density $n(z)$ and proper radius r_* .

First, we can determine the probability for a line of sight to intersect such a galaxy with redshift lying between z and $z+dz$. This probability is proportional to the surface area of the galaxy, to the density, and to the distance light propagated in this redshift interval,

$$dP = \pi r_*^2 n(z) D_{H_0} \frac{dz}{(1+z)E(z)}. \quad (3.87)$$

The probability to intersect an object with redshift smaller than z , called the *optical depth*, is then given by

$$\tau(z) = \pi r_*^2 D_{H_0} \int_0^z n(z) \frac{dz}{(1+z)E(z)}. \quad (3.88)$$

In particular, for galaxies diluted by the cosmic expansion, $n(z) = n_0(1+z)^3$, in a Universe with $\Omega = 1$ and $\Omega_\Lambda = 0$ we find

$$\tau(z) = \frac{2}{3} \pi r_*^2 D_{H_0} n_0 (1+z)^{3/2}. \quad (3.89)$$

If $n_0 \sim 0.02 h^3 \text{Mpc}^{-3}$ and $r_* \sim 10 h^{-1} \text{kpc}$, $\tau \sim 4\%$ at $z = 1$ and 100% at $z \sim 20$.

The proper volume contained in a solid angle $d\Omega^2$ and between the redshifts z and $z+dz$ is the product of the area $D_A^2 d\Omega^2$ and the length $D_{H_0} dz / (1+z)E(z)$, so that

$$\frac{dV}{d\Omega^2 dz} = D_{H_0} \frac{f_K^2[\chi(z)]}{(1+z)^3 E(z)}. \quad (3.90)$$

The number of galaxies in a solid angle $d\Omega^2$ and in a redshift band dz is then

$$\frac{dN}{d\Omega^2 dz} = D_{H_0} n(z) \frac{f_K^2[\chi(z)]}{(1+z)^3 E(z)}. \quad (3.91)$$

Galaxies number counts are among the oldest cosmological tests (see Ref. [1]).

3.4 Behaviour at small redshifts

3.4.1 Deceleration parameter

We can expand the scale factor around its value today as

$$a(t) = a_0 \left[1 + H_0(t - t_0) - \frac{1}{2} q_0 H_0^2 (t - t_0)^2 + \dots \right], \quad (3.92)$$

with $0 < t_0 - t \ll t_0$. The factor q_0 is the *deceleration parameter*⁷ and is defined by

$$q_0 \equiv - \left. \frac{\ddot{a}}{aH^2} \right|_{t=t_0}, \quad (3.93)$$

so that the Universe accelerates if $q_0 < 0$ and decelerates otherwise. Using the Friedmann equations to express \ddot{a}_0/a_0 , we easily find that

$$q_0 = \frac{1}{2} \Omega_{m0} - \Omega_{\Lambda 0},$$

for a Λ CDM model, since the contribution from radiation is negligible today. For an arbitrary matter composition, the deceleration parameter is given by

$$q_0 = \frac{1}{2} \sum_x (1 + 3w_x) \Omega_{x0}. \quad (3.94)$$

Thus, in order to have $q_0 < 0$, there should be at least one cosmic fluid (dominating today) with equation of state $w_x < -1/3$. We stress that the expansion (3.92) only relies on the hypothesis that the Universe is described by a Friedmann–Lemaître space-time. In particular, it makes no assumptions on the matter content of the Universe, contrary to (3.94).

3.4.2 Expression of the time and distances

3.4.2.1 Small redshift expansions

At small redshifts, we can expand the function $E(z)$ as

$$E(z) = 1 + \frac{1}{2} \left[\sum_x 3(1 + w_x) \Omega_{x0} + 2\Omega_{K0} \right] z + \mathcal{O}(z^2) = 1 + (q_0 + 1)z + \mathcal{O}(z^2),$$

neglecting the contribution of radiation. Thus, we conclude that the look-back time is

$$\Delta t(z) = t_{H_0} \left[1 - \frac{1}{2}(q_0 + 2)z \right] z + \mathcal{O}(z^3). \quad (3.95)$$

The radial distance takes the form

⁷The choice of sign is historical. At the time q_0 was introduced it was thought that the expansion of the Universe was decelerating so that $q_0 > 0$ was expected. Chapter 4 details why we now think that $q_0 < 0$.

$$a_0 \chi(z) = D_{H_0} \left[1 - \frac{1}{2}(q_0 + 1)z \right] z + \mathcal{O}(z^3), \quad (3.96)$$

the comoving angular distance,

$$R_{\text{ang}}(z) = \frac{D_{H_0}}{a_0} \left[1 - \frac{1}{2}(q_0 + 1)z \right] z + \mathcal{O}(z^3), \quad (3.97)$$

the angular distance,

$$D_A(z) = D_{H_0} \left[1 - \frac{1}{2}(q_0 + 3)z \right] z + \mathcal{O}(z^3), \quad (3.98)$$

and the luminosity distance,

$$D_L(z) = D_{H_0} \left[1 - \frac{1}{2}(q_0 - 1)z \right] z + \mathcal{O}(z^3). \quad (3.99)$$

Locally ($z \ll 1$) all the distances reduce to $D_{H_0}z$ and the deviations only depend on the value of the deceleration parameter. A distance-redshift diagram will thus be sensitive to H_0 in its region $z \ll 1$ (slope of the diagram), to q_0 mainly in its intermediate region ($z \sim 1$) and to the full set of cosmological parameters for larger z . This illustrates the evolution of the degeneracy of the function $E(z)$ represented in Fig. 3.7.

3.4.2.2 Time drift of the redshift

Suppose we observe the same galaxy at t_0 and $t_0 + \delta t_0$. There will be a change in its redshift given by

$$1 + z + \delta z = \frac{a(\eta_0 + \delta\eta_0)}{a(\eta + \delta\eta)} = (1 + z) \frac{1 + H_0\delta\eta_0}{1 + H\delta\eta}.$$

Now, null geodesics are straight lines in conformal coordinates so that $\delta\eta = \delta\eta_0 = \delta t_0/a_0 = \delta t/a$ from which we deduce that $\delta z/(1 + z) = [H_0 - H]\delta\eta_0$ to first order in $\delta\eta_0$ and thus

$$\frac{\delta z}{\delta t_0} = H_0 [(1 + z) - E(z)], \quad (3.100)$$

where E is defined by (3.39). This gives the time evolution of the redshift of any object at redshift z as a function of the cosmological parameters.

At low redshift, it reduces to $\delta z/\delta t = -q_0 H_0 z + \mathcal{O}(z^2)$. Indeed, this drift is extremely small and of the order of $\delta z/\delta t \sim 10^{-8}/\text{century}$. However, it gives a direct information on $E(z)$ and thus on the cosmological parameters (see Refs. [17, 18]). Note that the redshifts are free of the evolutionary properties of the sources, and that a direct detection of their drift would represent a new (and yet never exploited) and direct approach in observational cosmology to constrain the evolution history of the Universe, alternative to those based on luminosity or diameter distances. It can be of importance in the study of the dark-energy problem (see Chapter 12).

3.5 Horizons

An horizon is a frontier that separates observable events from non-observable ones. Two very different concepts of horizon have to be distinguished in cosmology (see Refs. [10,19]). The different quantities defined in this section are represented in Figs. 3.11 and 3.12.

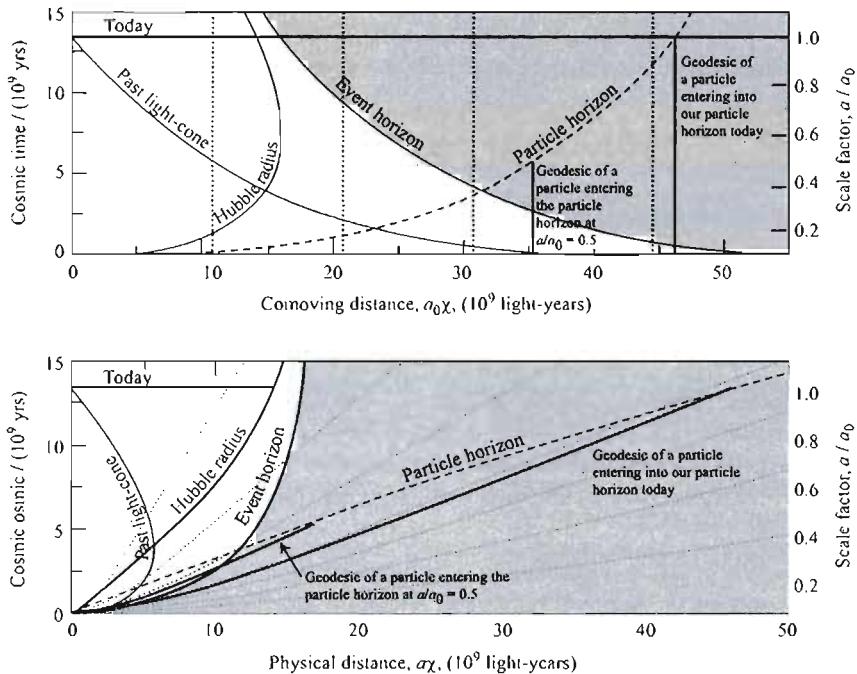


Fig. 3.11 The particle horizon at a given time (3.102) is defined at different moments as the geodesic of the most distant observable comoving particles at this moment (plain vertical lines). This surface can be visualized as the intersection of the particle horizon (3.103) with a constant-time hypersurface. The equation of this surface is given by $\chi = \phi(t)$ in comoving coordinates (top) and by $\chi = \sigma(t)$ in physical coordinates. This diagram represents a Universe with the following cosmological parameters $(\Omega_{m0}, \Omega_{\Lambda0}) = (0.3, 0.7)$ and $h = 0.7$. From Ref. [20].

3.5.1 Event horizon

For an observer O , the event horizon is the hypersurface that divides all events into two families: the ones that have been, are, or will be observable by O and the ones that are for ever outside the observational perimeter of O .

A necessary and sufficient condition for the existence of an event horizon is that the integral

$$\int^{\infty} \frac{dt}{a(t)}$$

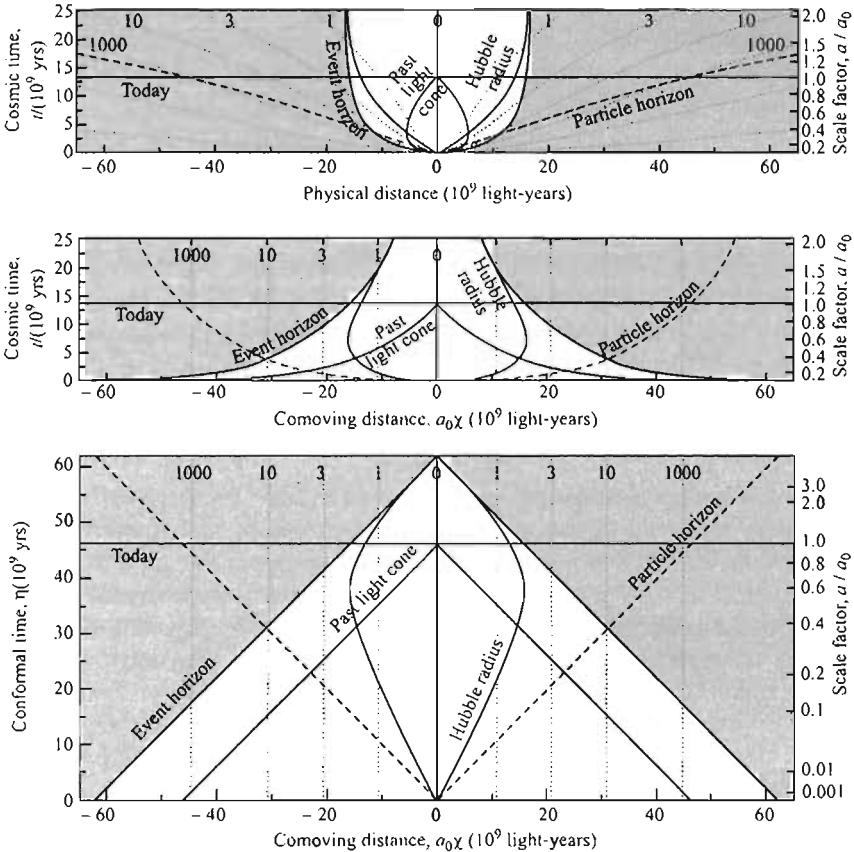


Fig. 3.12 Universe diagrams for a Friedmann–Lemaître space with parameters $\Omega_{m0} = 0.3$, $\Omega_{\Lambda0} = 0.7$ and $h = 0.7$. The first two diagrams represent, respectively, the physical distance, $a(t)\chi$, and the comoving distance, χ , in terms of the cosmic time, t (left scale) or in terms of the scale factor, normalized to 1 today (right scale). The last diagram represents the comoving distance in terms of the conformal time, η . The dotted lines represent the worldlines of comoving observers and our worldline is the central vertical line. Above each of these lines is indicated the redshift at which a galaxy on this worldline becomes visible for the central observer. We represent the past light cone for the present central observer and the event horizon corresponding to a similar light cone originated from time-like infinity (plain bold line). We also show the Hubble sphere, defined by (3.63) (plain light line) and the particle horizon, defined by (3.103) (dashed line). From Ref. [20].

is convergent. If this integral converges, then at any time t_0 , there exists a worldline

$$\chi = \chi_0 = \int_{t_0}^{\infty} \frac{dt}{a(t)},$$

such that the photon emitted at t_0 in χ_0 towards the origin, reaches $\chi = 0$ at $t = +\infty$.

Any photon emitted at t_0 in $\chi > \chi_0$ never reaches the origin and any photon emitted at t_0 in $\chi < \chi_0$ reaches the origin in a finite time.

The de Sitter space, for which $a = \exp(Ht)$, has such an event horizon, since

$$\chi_0 = \int_{t_0}^{\infty} \frac{dt}{\exp Ht} = \frac{e^{-Ht_0}}{H} < \infty.$$

As a second example, let us consider Universes with scale-factor evolution of the form $a(t) = t^n$. The integral

$$\int^{\infty} t^{-n} dt$$

converges if and only if $n > 1$. Recalling that for a flat Universe ($K = 0$) n is related to the parameter w of the equation of state by $3n = 2/(1+w)$, we find that an event horizon exists if

$$w < -\frac{1}{3} \iff \rho + 3P < 0, \quad (3.101)$$

i.e. if the strong energy condition is violated.

3.5.2 Particle horizon

For an observer O at t_0 , the particle horizon is the surface of the hypersurface $t = t_0$ dividing all particles of the Universe into two non-empty families: the ones that have already been observed by O at the time t_0 and the ones that have not. For each time t_0 , it is thus the intersection between the geodesics of the most distant comoving particles that can be observed, with the hypersurface $t = t_0$. Hence, it is a two-dimensional space-like surface, which reduces to a sphere of centre O for an isotropic and homogeneous Universe.

A necessary and sufficient condition for the existence of a particle horizon is that the following integral be convergent:

$$\int_0^{\infty} \frac{dt}{a(t)} \quad \text{or} \quad \int_{-\infty}^{t_0} \frac{dt}{a(t)},$$

depending on whether or not $a(t)$ continues for negative values of t .

Indeed, at any time t_0 , any particle such that $\chi > \int_0^{t_0} a^{-1} dt$ has not yet been observed by an observer O at the origin. The two-dimensional surface

$$\chi = \phi(t_0) = \int_0^{t_0} \frac{dt}{a(t)} \quad (3.102)$$

defines the particle horizon at a given time t_0 and thus divides all particles into two sub-families: the ones that have been observed at t_0 or before t_0 [$\chi \leq \phi(t_0)$] and the ones that have not yet been observed [$\chi > \phi(t_0)$]. This surface is called the particle horizon of O at t_0 . It can be seen as the section at $t = t_0$ of the space-time surface $\sigma = \phi(t)$. We thus define the particle horizon as the hypersurface

$$\chi = \sigma = \phi(t). \quad (3.103)$$

The hypersurface (3.103) is the future light cone emitted from the position of the observer at $t = 0$. Since $a(t)$ is a positive function, when $\phi(t)$ exists, it is an increasing

function of t so that as time elapses, more and more particles are visible from O . In conformal time, the particle horizon is a cone and the particle horizon at t_0 is a sphere.

If $\phi(t)$ is finite when t tends to infinity, then the Universe has an event horizon since only the particles with $x \leq \phi(\infty)$ will be accessible to the observer O . This is represented in Fig. 3.11 where the vertical lines represent the particle horizons at various times, whereas the dashed line represents the surface defined by $x = \phi(t)$.

As an example, let us consider Universes with scale-factor evolution of the form $a(t) = t^n$. The integral

$$\int_0^t t^{-n} dt$$

converges if and only if $n < 1$. Since for a flat Universe ($K = 0$), $3n = 2/(1+w)$, it follows that there exists a particle horizon if

$$w > -\frac{1}{3} \iff \rho + 3P > 0. \quad (3.104)$$

From the conditions (3.101) and (3.104), we thus conclude that a Universe containing a single fluid with constant equation of state can have either an event horizon or particle horizon, but not both at the same time. The two types of horizons are thus mutually exclusive in this particular case.

The equation (3.102) gives the comoving radius of the particle horizon. Its physical diameter at a time t_2 for an event that occurred at $t_1 < t_2$ is thus

$$D_{\text{p.h.}}(t_1, t_2) = 2a(t_2) \int_{t_1}^{t_2} \frac{dt}{a(t)}. \quad (3.105)$$

It can be checked that if t_1 and t_2 are two events from an era dominated by a fluid with constant equation of state w , then

$$D_{\text{p.h.}}(t_1, t_2) = \frac{6(1+w)}{1+3w} t_2 \left[1 - \left(\frac{t_1}{t_2} \right)^{(1+3w)/(3+3w)} \right]. \quad (3.106)$$

If $t_1 \ll t_2$ then the horizon diameter is proportional to the Hubble radius (3.63) at the time t_2

$$D_{\text{p.h.}}(t_1, t_2) \simeq \frac{4}{1+3w} D_H(t_2). \quad (3.107)$$

The particle horizon is at the origin of the ‘horizon problem’ of the standard Big-Bang model (see Section 4.5.2 of Chapter 4).

3.5.3 Global properties

To study the global structures, and in particular at infinity, of the cosmological space-times, we shall use the representation introduced by Penrose (see Ref. [10] for a general study). It is based on the idea of constructing, for any manifold \mathcal{M} with metric $g_{\mu\nu}$, another manifold $\tilde{\mathcal{M}}$ with a boundary \mathcal{J} and metric $\tilde{g}_{\mu\nu} = W g_{\mu\nu}$ such that \mathcal{M} is conformal to the interior of $\tilde{\mathcal{M}}$, and so that the ‘infinity’ of \mathcal{M} is represented by the ‘finite’ hypersurface \mathcal{J} . The last property implies that W vanishes on \mathcal{J} . All asymptotic properties of \mathcal{M} can be investigated by studying \mathcal{J} (see Ref. [21]).

3.5.3.1 The Minkowski space-time as an example

The Minkowski metric (1.10) in spherical coordinates can be written in terms of the advanced and retarded null coordinates, respectively, defined by $v = t+r$ and $u = t-r$ as

$$ds^2 = -dvdu + \frac{1}{4}(v-u)^2 d\Omega^2,$$

where $v \geq u$ and v and u range from $-\infty$ to $+\infty$. No terms of the form dv^2 or du^2 appear because the surfaces of constant v (or constant u) are null surfaces.

We can make these coordinates vary in a finite interval by introducing

$$\tan V = v, \quad \tan U = u,$$

so that $-\pi/2 < U \leq V < \pi/2$. Then, introducing the coordinates T and R by

$$T = U + V, \quad R = V - U,$$

with

$$-\pi < T+R < \pi, \quad -\pi < T-R < \pi, \quad R \geq 0, \quad (3.108)$$

the Minkowski metric turns out to be conformal to the metric \bar{g} given by

$$d\bar{s}^2 = -dT^2 + dR^2 + \sin^2 R d\Omega^2,$$

with conformal factor $W = \{2\sin[(R+T)/2]\sin[(T-R)/2]\}^{-2}$. The Minkowski space-time is thus conformal to the region (3.108) of the Einstein static space-time (a cylinder $S^3 \times \mathbb{R}$ that represents a static space-time with spherical spatial sections). The boundary of this region therefore represents the conformal structure of infinity of the Minkowski space-time.

This boundary can be decomposed into

- two 3-dimensional null hypersurfaces, \mathcal{J}^+ and \mathcal{J}^- defined by

$$\mathcal{J}^+ = \left\{ V = \frac{\pi}{2}, \quad |U| < \frac{\pi}{2} \right\}, \quad \mathcal{J}^- = \left\{ U = \frac{\pi}{2}, \quad |V| < \frac{\pi}{2} \right\}, \quad (3.109)$$

or equivalently by $T = \pm(\pi - R)$ with $R \in [0, \pi]$. The image of a null geodesic originates on \mathcal{J}^- and terminates at \mathcal{J}^+ , which represent future and past null-infinity.

- two points i^+ and i^- defined by

$$i^\pm : \quad U = V = \pm \frac{\pi}{2}, \quad (3.110)$$

or equivalently by $R = 0$ and $T = \pm\pi$. The image of a time-like geodesic originates at i^- and terminates at i^+ . They represent future and past time-like infinity, that is, respectively, the start- and end-point of all time-like geodesics.

- one point i^0 defined by

$$i^0 : \quad U = -V = -\frac{\pi}{2}, \quad (3.111)$$

or equivalently by $R = \pi$ and $T = 0$. It is the start- and end-point of all space-like geodesics so that it represents spatial infinity.

The fact that i^\pm and i^0 are single points follows from the fact that $\sin R = 0$. These are coordinate singularities of the same type as the one encountered at the origin of polar coordinates. The manifold \mathcal{M} is regular at these points. \mathcal{J}^- is a future null cone with vertex i^- and it refocuses to a point i^0 that is spatially diametrically opposite to i^- . The future null cone of i^0 is \mathcal{J}^+ , which refocuses at i^+ .

Note that the boundary is determined by the space-time and is unique but that the conformal extension space-time (here the Einstein static space-time) is not fixed by the original metric and is not unique since another conformal transformation could have been chosen.

For any spherically symmetric space-time, the region (3.108) can be represented in the (T, R) plane by ignoring the angular coordinates so that each point represents a sphere S^2 . In this representation the Minkowski space-time is a square lozenge (see Fig. 3.13). Null geodesics are represented by straight lines at $\pm 45^\circ$ deg running from \mathcal{J}^- to \mathcal{J}^+ .

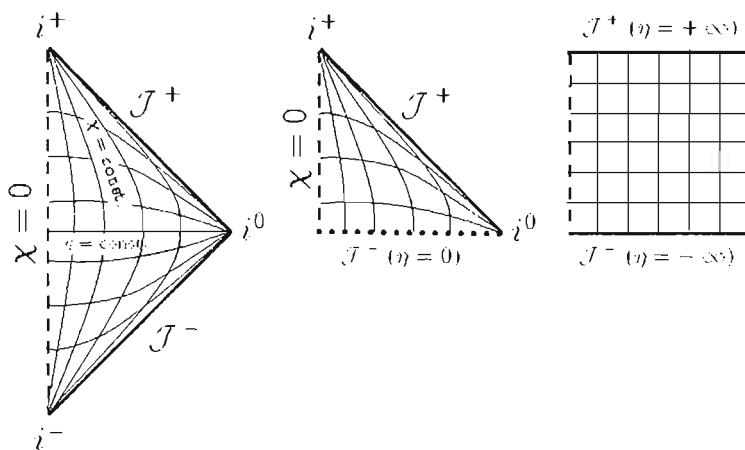


Fig. 3.13 Conformal diagrams of the Minkowski space-time (left) and Friedmann–Lemaître space-times with Euclidean spatial sections with $\Lambda = 0$ and $P > 0$ (middle) and de Sitter space (right).

3.5.3.2 Friedmann–Lemaître space-times with $K = 0$

When written in terms of the conformal time, Friedmann–Lemaître space-times with Euclidean spatial sections are conformal to Minkowski. It follows that they map on a part of region (3.108) representing Minkowski space-time in the Einstein static Universe. The actual region is determined by the range of variation of η . For $\Lambda = 0$ and $P > 0$, $0 < \eta < \infty$ so that it is conformal to the upper half of the Minkowski diamond defined by $T > 0$ (see Fig. 3.13) with a singularity boundary, $T = 0$.

3.5.3.3 de Sitter space-time

The metric of the de Sitter space-time is given by (3.44) with t varying in \mathbb{R} . It has spherical spatial sections and is conformal to the whole Einstein static space-time. To study its infinity we define the coordinates

$$R = \chi, \quad T = 2 \arctan(e^{Ht}), \quad (3.112)$$

with $0 < R < \pi$ and $0 < T < \pi$. It is then conformal to a square region of the Einstein static space with conformal factor $W = H^{-2} \cosh^2(Ht)$ (see Fig. 3.13).

We see that this conformal diagram has a space-like infinity both for time-like and null geodesics, in the future and in the past.

3.5.3.4 Friedmann–Lemaître space-times with $K = \pm 1$

Friedmann–Lemaître space-times with spherical spatial sections ($K = +1$) are conformal to the Einstein static space-time (with $\eta \rightarrow T$ and $\chi \rightarrow R$). They are thus mapped into the part of this space-time determined by the allowed values for η .

There are three general possibilities. When $\Lambda = 0$, η varies from 0 to π if $P = 0$ and from 0 to $\alpha < \pi$ when $P > 0$. The Friedmann–Lemaître space-time is thus conformal to a square region of the Einstein static space-time so that both \mathcal{J}^+ and \mathcal{J}^- are space-like and represent two singularities. When $\Lambda \neq 0$, η can vary from 0 to ∞ for the hesitating Universes (such as examples N and M in Fig. 3.3) or from $-\infty$ to ∞ for the bouncing Universes (such as examples G and H in Fig. 3.3). The Friedmann–Lemaître space-time is thus conformal to either half of or the entire Einstein static space-time. As in the case of the de Sitter space, the conformal region will be a square of the Einstein static space and \mathcal{J}^+ and \mathcal{J}^- are also space-like but do not necessarily represent a singularity.

In the case of Friedmann–Lemaître space-times with hyperbolic spatial sections ($K = -1$), the metric can be shown to be conformal to part of the region (3.108) by means of the coordinate transformation

$$\begin{aligned} T &= \arctan \left[\tanh \left(\frac{\eta + \chi}{2} \right) \right] + \arctan \left[\tanh \left(\frac{\eta - \chi}{2} \right) \right], \\ R &= \arctan \left[\tanh \left(\frac{\eta + \chi}{2} \right) \right] - \arctan \left[\tanh \left(\frac{\eta - \chi}{2} \right) \right]. \end{aligned}$$

Again, the exact shape of this region depends on the matter content (equation of state and Λ).

We see in these examples that some parts of the boundary correspond to the Big-Bang singularity $a = 0$. When $P > 0$ and $\Lambda = 0$, the initial singularity is space-like, which corresponds to the existence of a particle horizon. When $K = +1$ and $\Lambda = 0$, the future boundary is space-like, which signals the existence of event horizons for the fundamental observers.

For most physically interesting space-times, \mathcal{J}^+ and \mathcal{J}^- are either space-like or null, which is closely related to the existence of one of the two types of horizon described above.

If \mathcal{J}^- is space-like, the worldlines of the fundamental observers do not all meet on \mathcal{J}^- at the same point. For any particular observer and event on its worldline close to \mathcal{J}^- , the past light cone of this event will not intercept all the particles in the Universe before it reaches \mathcal{J}^- and there will be a particle horizon. If \mathcal{J}^- is null, it is expected that all the worldlines of the fundamental observers pass through the vertex i^- so that the past light cone of a point P will intercept all the worldlines and there is no particle horizon.

If \mathcal{J}^+ is space-like, the worldline of any fundamental terminates on a point O of \mathcal{J}^+ and the past light cone of O divides the Universe into events that can be seen by the observer and events that he can never see; there is an event horizon. If \mathcal{J}^+ is null, all the worldlines will pass through the vertex i^+ , so that $O = i^+$ and its past light cone is \mathcal{J}^+ and there is no event horizon.

In conclusion,

$$\begin{aligned}\mathcal{J}^- \text{ space-like} &\iff \text{existence of a particle horizon,} \\ \mathcal{J}^+ \text{ space-like} &\iff \text{existence of an event horizon.}\end{aligned}$$

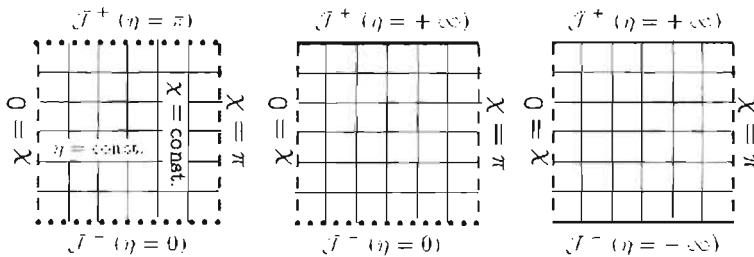


Fig. 3.14 Conformal diagrams of the Friedmann–Lemaître space-times with spherical spatial sections with $P > 0$ for $\Lambda = 0$ (left), and $\Lambda \neq 0$: hesitating case (middle) and bouncing case (right). The dotted line represents a singularity. The three diagrams differ only by the nature of \mathcal{J}^\pm , even though they are all space-like surfaces.

3.6 Beyond the cosmological principle

The Friedmann–Lemaître space-times have spatial sections with constant curvature, i.e. three-dimensional homogeneous and isotropic surfaces. These symmetries are characterised by six Killing vectors (Chapter 1) that completely fix the geometry of the spatial metric, up to a number, K .

Reciprocally, we can construct some solutions by restraining the number of symmetries and by imposing the structure of their Lie algebra. Numerous solutions of Einstein’s equations are known in this case and have been classified [22–24].

We give a general description of the classification of cosmological space-times according to their symmetries and then we present two cosmological solutions in order to illustrate the richness that this provides.

3.6.1 Classifying space-times

Even though the previous discussion focused on a particular class of cosmological solutions with maximally symmetric spatial sections, most solutions of Einstein's equations have no symmetry at all. This simplification was motivated by the cosmological principle but it is important to have an idea of exact solutions with less symmetry, in particular to construct counterexamples to some conclusions or test their robustness.

3.6.1.1 Symmetries and Killing vectors

Symmetries of a given space are transformations of this space into itself leaving the metric unchanged as well as all its geometrical properties. As explained in Chapter 1, a continuous symmetry can be characterized by a Killing vector field, ξ , satisfying

$$\mathcal{L}_\xi g_{\mu\nu} = \nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu = 0.$$

The group of isometries of a Riemannian manifold forms a Lie group (that is a group that is also a smooth manifold with a group operation xy^{-1} that is a smooth map into the manifold itself; see Chapter 2). The set of Killing vector fields generators of these isometries forms the associated *Lie algebra*: it is a vector space, since the linear combination of any two Killing vectors is also a Killing vector, of finite dimension r since the maximum number of Killing vectors for a space of dimension n is $n(n+1)/2$, reached for maximally symmetric spaces. The product operation is defined by the commutator of two Killing vectors

$$[\xi_a, \xi_b] \equiv \mathcal{L}_{\xi_a} \xi_b = -[\xi_b, \xi_a] \quad (3.113)$$

of coordinates $[\xi_a, \xi_b]^\alpha = (\partial_\beta \xi_b^\alpha) \xi_a^\beta - (\partial_\beta \xi_a^\alpha) \xi_b^\beta$ and it satisfies the Jacobi identity, $[\xi_a, [\xi_b, \xi_c]] + [\xi_b, [\xi_c, \xi_a]] + [\xi_c, [\xi_a, \xi_b]] = 0$.

It is easily checked that if ξ_a and ξ_b are two Killing vectors then $[\xi_a, \xi_b]$ is also a Killing vector (since $\mathcal{L}_{[\xi_a, \xi_b]} g_{\mu\nu} = \partial_\mu [\xi_a, \xi_b]_\nu + \partial_\nu [\xi_b, \xi_a]_\mu = 0$). Considering a basis $\{\xi_a\}_{a=1..r}$ of this algebra, any Killing vector can be decomposed on this basis as a linear combination with constant coefficients. It follows that

$$[\xi_a, \xi_b] = C_{ab}^c \xi_c, \quad (3.114)$$

where C_{ab}^c are the *structure constants* of the Lie algebra. They are antisymmetric, $C_{ab}^c = -C_{ba}^c$, and must satisfy, from the Jacobi identity $C_{e[b}^a C_{cd]}^e = 0$. These relations are integrability conditions that ensure that the Lie algebra is defined in a consistent way.

3.6.1.2 Definitions

The action of a group of isometries defines orbits in the space where it acts and the dimensionality of these orbits characterizes the symmetries.

The orbit of a point P is the set of all points into which P is transformed by the action of the isometries of the space. The orbits are thus invariant sets. An *invariant variety* is a set that is globally left invariant by the action of the group of isometries. It will be larger than all the orbits it contains. The orbits are the smallest invariant subspaces (for all the isometries in the group). An orbit that reduces to a single point is a *fixed point*. It is a point where all Killing vectors vanish. We distinguish *general points* where the dimension of Killing Lie algebra takes the value it has almost everywhere from *special points*, such as the fixed points, where the dimension is lower.

The group of isometries is *transitive on a space S* if it can move any point of S into any other point of S . Orbits are the largest spaces through each point on which the group is transitive and can be called *surfaces of transitivity*. Their dimension, s , has to be smaller than or equal to the dimension of the group of translation,

$$s = \dim(\text{surface of transitivity}) \leq n.$$

The *isotropy group*, at a point P , is the subgroup of isometries letting that particular point be fixed. It is generated by the Killing vectors that vanish at P . Its dimension q is smaller than the dimension of the group of rotations, $n(n - 1)/2$,

$$q = \dim(\text{isotropy group}) \leq \frac{1}{2}n(n - 1).$$

The dimension of the group of isometries, r , is the sum of the dimensions of the isotropy group and of the surface of transitivity. It follows that r has to be smaller than the maximum dimension of the Killing Lie algebra, that is $n(n + 1)/2$, that is for a space of constant curvature.

$$r = \dim(\text{group of isometries}) = s + q \leq \frac{1}{2}n(n + 1).$$

The group of isometries of a space is thus characterized by two of the three numbers (s, q, r) and by the structure constants of the group (C_{ab}^c).

3.6.1.3 Classification of cosmological models

For a $n = 4$ dimensional space-times, $r \leq 10$ and s can run from 0 to 4. The dimension of the group of isometries being a global property of the space, r must not change over space. Indeed, the real Universe has no symmetry ($r = 0$).

For cosmological spaces, we assume that $\rho + P > 0$, which implies that $q \in \{3, 1, 0\}$ because the congruence of the worldline of the fundamental observers is invariant, so that the isotropy group at each point must be a subgroup acting orthogonally to u^μ and it can be shown that there is no subgroup of $O(3)$ of dimension 2. Furthermore, q can vary over space but not over an orbit and it can be greater at special points where s is smaller. The space will be said to be *isotropic* for $q = 3$ (this is the case of the Friedmann–Lemaître spaces), *locally invariant under rotation* for $q = 1$ (there is a preferred spatial direction) and *anisotropic* for $q = 0$.

Table 3.2 summarizes all the possibilities according to q and s . For more details, see Refs. [7, 22–24].

Table 3.2 Classification of cosmological models according to their isotropy and homogeneity properties. We recall that $\tau = q + s \leq 10$.

| | $s = 4$ space-time homogeneous | $s = 3$ spatially homogeneous | $s = 2$ inhomogeneous |
|---------|--|----------------------------------|--------------------------|
| $q = 6$ | Minkowski, de Sitter anti-de Sitter | none | none |
| $q = 3$ | Einstein static | Friedmann–Lemaître | none |
| $q = 1$ | | Bianchi Kantowski–Sachs | Lemaître–Tolman–Bondi |
| $q = 0$ | | Bianchi | |

3.6.2 Universe with non-homogeneous spatial sections

3.6.2.1 Lemaître–Tolman–Bondi space-time

Among all the non-homogeneous spaces, the Lemaître–Tolman–Bondi space-time is the simplest. It enjoys a spherical symmetry, so that $s = 2$, $q = 1$ and $r = 3$ (but note that the origin is a special point with $q = 3$ and $s = 0$). Its metric takes the general form

$$ds^2 = -dt^2 + S^2(t, r)dr^2 + R^2(t, r)d\Omega^2. \quad (3.115)$$

The Einstein equations for a fluid of dust ($P = 0$) reduce to

$$S^2(r, t) = \frac{(R')^2}{1 - K(r)r^2} \quad (3.116)$$

$$\dot{R}^2 = 2\frac{M(r)}{R} - K(r)r^2 \quad (3.117)$$

$$4\pi G_N \rho(r, t) = \frac{M'(r)}{R^2 R'}, \quad (3.118)$$

where a prime and a dot represent the derivatives with respect to r and t , respectively. $M(r)$ and $K(r)$ are two integration functions. The Friedmann–Lemaître space-times correspond to the special case $R = a(t)r$ and $K(r) = K$. These equations can be integrated to give the solution

$$R(\eta, r) = \frac{M(r)}{2\mathcal{E}(r)} \frac{d\Phi}{d\eta}, \quad t(\eta, r) - t_0(r) = \frac{M(r)}{(2\mathcal{E})^{3/2}} \Phi(\eta) \quad (3.119)$$

where $2\mathcal{E} = (2E, 1, -2E)$, $\Phi = (\sinh \eta - \eta, \eta^3/6, \eta - \sin \eta)$ for, respectively, positive, zero, or negative $E = -K(r)r^2/2$.

Thus, the solutions depend on three arbitrary functions. $M(r)$ is the gravitational mass contained in the ball of radius r , $t_0(r)$ is the local time at which $R = 0$, i.e. the time of the Big Bang, and $E(r)$ determines if the spatial section is locally spherical, Euclidean or hyperbolic. The curvature of the spatial section is thus no longer uniform and can depend on the distance to the centre.

This solution has been used to study the structure of an inhomogeneous Big Bang, the effect of the cosmological principle on data interpretations, to model overdense regions and voids in an expanding space-time, or also to study the geometrical dipole that would be seen in the cosmic microwave background by an off-centre observer.

3.6.2.2 The Swiss-cheese model

This family of inhomogeneous space-times is constructed by cutting out spherical regions from a Friedmann–Lemaître Universe and filling the voids by a spherically symmetric solution of the Einstein equation, for example, a Schwarzschild solution. The two space-times are then glued along a 3-dimensional time-like hypersurface. This implies constraints on the two space-times since they have to obey junction conditions for the total space-time to be well behaved (see Section 8.4.5 of Chapter 8 for a description of these conditions).

For instance, it can be shown that it implies that the central mass of the Schwarzschild space-time, the density and scale factor of the Friedmann–Lemaître space-time and the radius of the spherical void must be related by $M_{\text{Schw}} = (4\pi/3)[\rho a^3 R^3]_{\text{FL}}$. Thus, we can construct a spherical void with an overdensity at the centre. The junction conditions impose that the over- and underdensities average to the mean density of the Friedmann–Lemaître space-time. It follows that the mass in the central void cannot have any effect on matter outside the void, and in particular on the expansion rate: its gravitational effect is screened.

3.6.3 Universe with homogeneous spatial sections

The Universe models with $s = 3$ are central in theoretical cosmology because they mathematically realize the cosmological principle (they have homogeneous spatial sections). We have described at length the case of isotropic space-times ($q = 3; r = 6$) in this chapter. Two classes of non-isotropic space-times can be constructed: (i) the family of solutions with a local rotation symmetry ($q = 1; r = 4$), which contains the family of Kantowski–Sachs Universes and some Bianchi Universes, and (ii) the family of anisotropic solutions ($q = 0; r = 3$), which contains the other Bianchi Universes.

3.6.3.1 Generalities on Bianchi Universes

To classify all spaces with $s = 3$, we need to determine the structure constants, C_{ab}^c , of the group that is simply transitive on space-like sections. The C_{ab}^c have 9 independent components and can thus be mapped onto the 3×3 matrix N_{ab} defined by

$$N^{cd} = \frac{1}{2} C_{ab}^c \varepsilon^{abd},$$

where ε^{abd} is the completely antisymmetric tensor. Splitting N_{cd} into its symmetric and antisymmetric parts as $N_{cd} = n^{cd} + \varepsilon^{cdb} A_b$, we obtain that the structure constants are

$$C_{ab}^c = \varepsilon_{dab} n^{cd} + \delta_b^c A_a - \delta_a^c A_b.$$

One can always choose the basis of the Lie algebra such that n^{ab} is diagonal [$n_{ab} = \text{diag}(n_1, n_2, n_3)$]. A rotation allows us to set $A_b = (a, 0, 0)$ while keeping n^{ab} diagonal. The Jacobi identity implies that $n^{ab} A_b = 0$, so that we have one of the three possibilities: (i) $a = 0$ and $n_1 \neq 0$, (ii) $a \neq 0$ and $n_1 = 0$, or (iii) $a = 0$ and $n_1 = 0$. By rescaling the Killing vectors, we can finally set the nonvanishing n_a to ± 1 but in general it is impossible to rescale both a and n_a . In conclusion

$$n_{ab} = \text{diag}(n_1, n_2, n_3), \quad A_b = (a, 0, 0) \text{ with } n_i \in \{0, \pm 1\} \text{ and } a n_1 = 0. \quad (3.120)$$

Table 3.3 gives the eleven groups of isometries that are distributed among the nine Bianchi types I to IX . They are divided into two classes: class A if $a = 0$ and class B otherwise. In some cases, these groups allow higher symmetry. For instance, Bianchi I and VII_0 can be isotropic (leading to $K = 0$ Friedmann-Lemaître space-times) as well as Bianchi IX (leading to $K = +1$ Friedmann-Lemaître space-times) and Bianchi V and Bianchi VII_a (leading to $K = -1$ Friedmann-Lemaître space-times).

Table 3.3 The structure constants of the 11 groups of isometries and the definition of the 9 associated Bianchi types.

| Class | A | | | | | | B | | | | | |
|-------|--------------|-----|------|---------|--------|------|--------|-----|-------|------|---------|--------|
| | Bianchi type | I | II | VII_0 | VI_0 | IX | $VIII$ | V | III | IV | VII_a | VI_a |
| a | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | a | a | |
| n_1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | |
| n_2 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | |
| n_3 | 0 | 1 | 1 | -1 | 1 | -1 | 0 | 1 | -1 | 1 | -1 | |

3.6.3.2 Example: Bianchi I Universe

The simplest Universe with anisotropic expansion is the Bianchi I Universe. It is a homogeneous space-time so that the projected gradient of any scalar function vanishes, $\bar{\nabla}_\nu f = 0$. This implies, in the notations of Chapter 1, that $a_\mu = 0$ (because $\bar{\nabla}_\mu P = 0$ and the vorticity is also vanishing, $\omega_{\mu\nu} = 0$). It follows that it is completely described by $\Theta(t)$, $\sigma(t)$, $\rho(t)$, and $P(t)$.

The Raychaudhuri equation (1.73) gives

$$\dot{\Theta} + \frac{1}{3}\Theta^2 + 2\sigma^2 = -4\pi G_N(\rho + 3P).$$

The spatial sections are assumed to be Euclidean so that ${}^{(3)}R = 0$, which implies that the generalized Friedmann equation (1.78) reduces to

$$\frac{1}{3}\Theta^2 - \sigma^2 = 8\pi G_N\rho.$$

The shear evolution equation is derived from the Gauss-Codazzi relation, using the fact that ${}^{(3)}R_{ab} = 0$, and is given by

$$u^c \nabla_c (\sigma_{ab}) + \Theta \sigma_{ab} = 0.$$

In terms of the cosmic time derivatives, this rewrites as

$$(\sigma_j^i)_{\cdot} + \Theta \sigma_j^i = 0.$$

In comoving coordinates, the metric of this space-time takes the form

$$ds^2 = -dt^2 + X^2(t)dx^2 + Y^2(t)dy^2 + Z^2(t)dz^2. \quad (3.121)$$

The average of the directional expansion rates is $S = (XYZ)^{1/3}$ and is related to Θ by $3\Theta = \dot{S}/S$. Factorizing S , the metric can be recast as

$$ds^2 = -dt^2 + S^2(t)\gamma_{ij}(t)dx^i dx^j, \quad (3.122)$$

and the metric on constant-time hypersurfaces can be decomposed as

$$\gamma_{ij} = e^{2\beta_i(t)}\delta_{ij}, \quad (3.123)$$

where the β_i have to satisfy $\sum \beta_i = 0$. $\dot{\gamma}_{ij}$ is traceless and related to the shear by

$$\sigma_{ij} = \frac{1}{2}\dot{\gamma}_{ij}.$$

The equation of evolution of the shear implies that

$$\dot{\sigma}_j^i = \frac{\Sigma_j^i}{S^3}, \quad (3.124)$$

where Σ_j^i a constant tensor. This implies that $\dot{\sigma}^2 \equiv \sigma_{ij}\sigma^{ij} = \Sigma^2/S^6$, where $\Sigma^2 = \Sigma_j^i\Sigma_i^j$, and thus that the generalized Friedmann equation (1.78) takes the form

$$3\frac{\dot{S}^2}{S^2} = \kappa\rho + \frac{1}{2}\frac{\Sigma^2}{S^6}. \quad (3.125)$$

For a fluid with equation of state $P = w\rho$, $\rho \propto S^{-3(w+1)}$. This implies that for any value of Σ , the shear always dominates the primordial evolution of the Universe. It can also be checked that the Raychaudhuri equation takes the form

$$\frac{\ddot{S}^2}{S^2} = -\frac{\kappa}{6}(\rho + 3P) - \frac{1}{3}\frac{\Sigma^2}{S^6}. \quad (3.126)$$

The evolution of the three scale factors can be shown to be obtained by setting

$$\beta_i(t) = B_i W(t), \quad (3.127)$$

with

$$W(t) = \int \frac{dt}{S^3}. \quad (3.128)$$

The coefficients B_i must satisfy the constraint

$$\sum B_i = 0, \quad \sum B_i^2 = \Sigma^2. \quad (3.129)$$

This relation is trivially verified with

$$B_i = \sqrt{\frac{2}{3}}\Sigma \sin \alpha_i, \quad \alpha_i = \alpha + \frac{2\pi}{3}i, \quad (3.130)$$

where α is a constant. For instance, if $w = 0$, we obtain the solution

$$S = \left(\frac{3}{4} M t^2 + \sqrt{\frac{3}{2}} \Sigma t \right)^{1/3}, \quad W = -\sqrt{\frac{2}{3}} \frac{1}{\Sigma} \ln \left(3M + 2\sqrt{6}\Sigma/t \right), \quad (3.131)$$

once setting $\kappa\rho = M/S^3$. At late times, the Universe becomes more and more isotropic and tends towards an Einstein-de Sitter space-time, but close to the Big Bang, the shear dominates, $S \rightarrow (\sqrt{3}\Sigma t)^{1/3}$ and

$$X(t) \rightarrow X_0 t^{(1+2\sin\alpha_1)/3}, \quad Y(t) \rightarrow Y_0 t^{(1+2\sin\alpha_2)/3}, \quad Z(t) \rightarrow Z_0 t^{(1+2\sin\alpha_3)/3}. \quad (3.132)$$

If $\alpha \neq \pi/2$, two of the powers are positive and the third one is negative. Similar behaviours close to the singularity will be discussed in Section 13.4.1 of Chapter 13.

As another example, consider the case of a pure cosmological constant. Then, the Friedmann equation can take the form

$$\frac{\dot{S}^2}{S^2} = H_*^2 \left(1 + \frac{S_*^6}{S^6} \right).$$

It follows that

$$S(t) = S_* |\sinh(3H_* t)|^{1/3}.$$

Again, at late time, it converges toward the de Sitter scale factor $S \sim \exp(H_* t)$ while at early time it behaves as $t^{1/3}$.

References

- [1] P.J.E. PEEBLES, *Principle of physical cosmology*, Princeton University Press, 1993.
- [2] S. WEINBERG, *Gravitation and cosmology: principles and applications of the general theory of relativity*, John Wiley and Sons, 1972.
- [3] J. RICH, *Fundamentals of cosmology*, Springer-Verlag, 2001.
- [4] E. HARISSON, *Cosmology*, Cambridge University Press, 1981.
- [5] B. RIDDEN, *Introduction to cosmology*, Addison Wesley, 2003.
- [6] R.C. TOLMAN, *Relativity, thermodynamics and cosmology*, Oxford University Press, 1934.
- [7] G.F.R. ELLIS and H. VAN ELST, ‘Cosmological models’, in *Theoretical and observational cosmology*, Kluwer, Dordrecht, 1999.
- [8] J.P. UZAN, C. CLARKSON and G.F.R. ELLIS, ‘Time drift of the cosmological redshift as a test of the Copernican principle’, *Phys. Rev. Lett.* **100**, 191303, 2008.
- [9] R. WALD, *Gravitation*, Chicago University Press, 1984.
- [10] S. HAWKING and G.F.R. ELLIS, *The large scale structure of space-time*, Cambridge University Press, 1973.
- [11] A.D. CHERNIN, D.I. NAGIRNER and S.V. STARIKOVA, ‘Growth rate of cosmological perturbations in standard model: explicit analytical solution’, *Astron. Astrophys.* **399**, 19, 2003.
- [12] J. EHRLERS and W. RINDLER, ‘A phase space representation of Friedmann–Lemaître Universes containing both dust and radiation and the inevitability of a Big Bang’, *Month. Not. R. Astron. Soc.* **238**, 503, 1989.
- [13] G.F.R. ELLIS and J. WAINWRIGHT (eds.), *The dynamical system approach to cosmology*, Cambridge University Press, 1996.
- [14] J.-P. UZAN and R. LEHOUCQ, ‘A dynamical study of the Friedmann equations’, *Eur. J. Phys.* **22**, 371, 2001.
- [15] G.F.R. ELLIS, ‘Relativistic cosmology’, in *Relativity and cosmology*, R.K. SACHS (ed.), Academic Press, 1971.
- [16] J.-P. UZAN, N. AGHANIM and Y. MELLIER, ‘The distance duality relation from X-ray and SZ observations of clusters’, *Phys. Rev. D* **70**, 083533, 2004.
- [17] M. DAVIES and L. MAY, ‘New observations of the radio absorption line in 3C 286, with potential application to direct measurement of cosmological deceleration’, *Astrophys. J.* **219**, 1, 1978.
- [18] A. LOEB, ‘Direct measurement of cosmological parameters from the cosmic deceleration of extragalactic objects’, *Astrophys. J. Lett.* **499**, L111, 1998.
- [19] W. RINDLER, ‘Visual horizons in world-models’, *Month. Not. R. Astron. Soc.* **116**, 622, 1953.

- [20] T.M. DAVIS and C.H. LINEWEAVER, *Expanding confusion: common misconceptions on cosmological horizons and the superluminal expansion of the Universe*, [astro-ph/0310808].
- [21] R. PENROSE, 'Conformal treatment of infinity', in *Relativity, groups and topology*, *Les Houches 1963*, pp. 561 C. DE WITT, and B. DE WITT (eds.), Gordon and Breach, 1964.
- [22] A. KRASINSKY, *Physics in an inhomogeneous Universe*, Cambridge University Press, 1996.
- [23] D. KRAMER, H. STEPHANI, M.A.H. MCCALLUM and E. HERTL, *Exact solutions of Einstein's field equations*, Cambridge University Press, 1980.
- [24] M.P. RYAN, *Homogeneous relativistic cosmologies*, Princeton University Press, 1975.

The standard Big-Bang model

This chapter is devoted to the description of the historical pillars on which the standard Big-Bang model stands. Starting from the cosmological solutions described in Chapter 3, we present the observations proving the expansion of the Universe and the various consequences of this expansion.

We start, in section 4.1, by describing the observational evidences of the expansion of the Universe from Hubble's observations to the most recent ones. This will lead us to discuss the value of the Hubble constant and its relation with the age of the Universe.

Elements of thermodynamics in an expanding space-time are summarised in section 4.2. We will deduce that the Universe has a thermal history. In particular, we will show how the departures from thermodynamic equilibrium play a central role.

A remarkable feature of the Big-Bang model comes from the proof by Alpher, Bethe and Gamow in 1948 [1] that light nuclei can be formed during *primordial nucleosynthesis*. This is discussed in section 4.3.

Gamow [2] also deduced the existence of a microwave background of photons, the temperature of which was calculated by Alpher and Herman [3] in 1948. Relying on an estimate of the helium abundance and of the baryon density, they concluded that this temperature should be around 5 K. Section 4.4 describes the properties of the isotropic component of this *microwave background radiation*.

These physical ingredients transform the (mathematical) cosmological solutions of Friedmann-Lemaître into the *Big-Bang model*. We will argue in section 4.5 that this model is an excellent standard model for cosmology, even if it suffers some problems, on which we will focus our attention in the following chapters of this book.

4.1 The Hubble diagram and the age of the Universe

If the Universe is expanding, galaxies must move away from one another (see Section 3.1.3; Chapter 3). The measurement of the systematic achromatic shift to longer wavelengths of the absorption or emission spectral lines gives access to the redshift z . For small redshifts, the recession velocity is given by $v/c \sim z$. As for distances, (3.99) shows that the luminosity and angular distances reduce to

$$D_A \sim D_L \sim \frac{c}{H_0} z \sim \frac{v}{H_0},$$

as long as $z \ll 1$. In this regime, the slope of the curve of v as a function of D directly gives the Hubble constant H_0 . For larger redshifts, this is no longer the case. First, the

angular distance and the luminosity distance are no longer identical, and in addition other cosmological parameters also interfere in the determination of these distances (see Chapter 3).

We describe the measurements for small redshifts in Section 4.1.1 and then those for supernovæ that extend up to redshifts of order 1 in Section 4.1.2. We finish by summarizing the methods allowing the age of the Universe to be determined in Section 4.1.3.

4.1.1 The Hubble constant

The Hubble diagram is the most direct proof of the expansion of the Universe. The underlying idea is very simple and comes from the independent measurements of the redshift and distance.

These measurements must be performed at distances large enough for the proper velocity of the galaxies to be negligible. Typically, one should measure velocities larger than $v \sim 10^4 \text{ km} \cdot \text{s}^{-1}$, which corresponds to a redshift of $z \sim 0.03$.

The original measurements of Hubble, lie respectively, up to velocities of order $v \sim 1000 \text{ km} \cdot \text{s}^{-1}$ and then up to velocities of order $v \sim 20000 \text{ km} \cdot \text{s}^{-1}$ (see Fig. 4.1). This analysis provided a value for H_0 that was 10 times larger than the one adopted today. For a galaxy with recession velocity $v \sim 10000 \text{ km} \cdot \text{s}^{-1}$, the effect of its proper velocity $v \sim 300 \text{ km} \cdot \text{s}^{-1}$ leads to an uncertainty of approximatively 3%. This uncertainty can be reduced by observing a large number of objects well distributed over the whole sky.

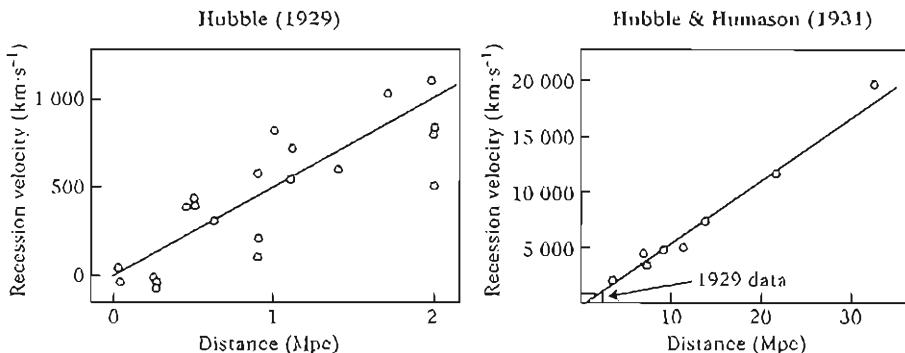


Fig. 4.1 Observations by Hubble in 1929 (left) and by Hubble and Humason in 1931 (right) proving the expansion of the Universe. The first measurement of H_0 included 18 galaxies for which both the distance and the redshift were determined. Hubble concluded that $H_0 = 500 \text{ km} \cdot \text{s}^{-1} \cdot \text{Mpc}^{-1}$ [4].

4.1.1.1 Astronomical methods

The difficulty in constructing a reliable scale of distances [5] has limited progress in the measurement of the Hubble constant. Thanks to the Hubble Space Telescope (HST), it is now possible to measure H_0 with a precision of about 10%. In astronomy, distances cannot be measured directly. The parallax method, which is purely geometrical, can

only be used to determine the distances of the closest stars in the Milky Way. The general method is then to measure the intrinsic luminosity of a class of objects for which it is either constant, or correlated to another physical parameter, whose value is thought to be independent of the distance, as a way to calibrate the relative distances. Various classes of objects have been used to this purpose (see Ref. [6] for a review).

1. *Cepheids*: cepheids are variable stars the external atmosphere of which pulses with a period between 2 and 100 days. These young and bright stars are very abundant in the close spiral and irregular galaxies. The pulsation mechanism is well understood theoretically and it has been established empirically that the pulsation period, which is independent of the distance, is correlated to the intrinsic luminosity of the star. The dispersion of the period-luminosity relation is of the order of 20% in luminosity. Since the luminosity scales as the inverse of the square of the distance, this implies a precision of 10% on the determination of the distance. For a sample of 25 cepheids, the statistic uncertainty drops to 2%. The Hubble diagram obtained from cepheids by the HST Key Project [7] is presented in Fig. 4.2.

The most precise calibration of the period-luminosity relation is performed on the Large Magellanic Cloud (LMC) and its distance has been determined with a precision of 10%.

At the moment, the detection of cepheids is limited to about 20 Mpc. To extend the scale, objects brighter than simple stars should be found. Their luminosity can be calibrated by identifying such objects in the nearby galaxies that also contains cepheids.

2. *Type Ia supernovæ*: type Ia supernovæ (SN Ia) are the most promising distance indicators for modern observational cosmology. The origin of SN Ia is the explosion of a white dwarf. Their luminosity is comparable to that of an entire galaxy, making them observable at distances of several hundreds Mpc. As will be shown later in Section 4.1.2, their luminosity curve shows a strong correlation between the characteristic evolution time and the maximal luminosity. The uncertainty of this relation is of the order of 12% on the luminosity, which translates into an uncertainty of around 6% in the determination of their distance.
3. *The Tully–Fisher relation*: the total luminosity of spiral galaxies is strongly correlated to the maximal rotation velocity of the galaxy. This relation, detailed in Chapter 7, reflects the fact that a brighter galaxy, and thus a more massive one, must rotate faster to compensate the gravitational attraction. The rotation velocity can be measured by spectroscopic observations from the Doppler effect. The Tully–Fisher relation has been measured for around a hundred galaxies and it can be shown empirically that it has a dispersion of the order of 30% in luminosity, which implies a precision of the order of 15% in the determination of the distance.
4. *Fundamental plane*: the intrinsic luminosity of elliptical galaxies is correlated to the velocity dispersion of their stars. This correlation is analogous to the Tully–Fisher relation. Elliptical galaxies occupy a ‘fundamental plane’ where their luminosity L is correlated to the surface brightness I_0 and to the velocity dispersion σ , by a relation of the form $L \propto I_0^{-0.7} \sigma^3$.

This relation can be understood by the fact that elliptical galaxies are roughly

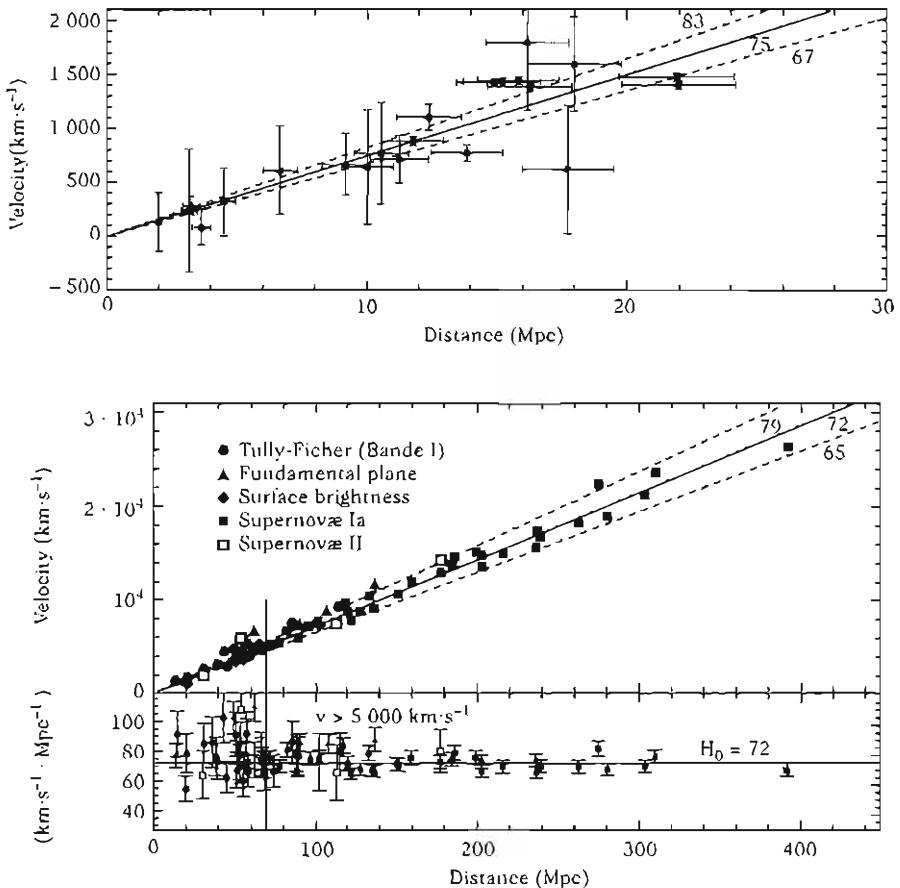


Fig. 4.2 (top): Hubble diagram for cepheids obtained by the HST Key Project collaboration [7]. Data are spread up to distances of around 20 Mpc, just as the measurement of Hubble and Humason (1931). (bottom): Different objects are used and thanks to type Ia supernovæ, the diagram can spread up to around 300 Mpc.

self-gravitating systems with a mass-to-luminosity ratio more or less constant. The virial theorem (see Section 7.2.2) then imposes $\sigma \propto M/r_0$. Moreover, the surface brightness is well described by $I(r) = I_0 \exp[-(r/r_0)^{1/4}]$, called the de Vaucouleur law. This can be integrated as $L \propto I_0 r_0^2$. Assuming that the mass-to-luminosity ratio is of the form $M/L \propto M^x$, we obtain that $L^{1+x} \propto \sigma^{4-4x} I_0^{x-1}$. The empirical law is recovered for $x \sim 0.25$.

The precision of this relation is of the order of 10–20%, which implies a precision of the order of 5–10% on the determination of the distance.

5. *Fluctuation of the surface brightness:* the resolution of stars with a CCD camera depends on the distance. The luminosity of each pixel is due to a given number of stars. One can show that the Poisson fluctuations from one pixel to another

depend on the distance of the galaxy. This method has a precision of the order of 8% and is applied to redshifts up to $z \sim 0.02$.

4.1.1.2 Summary of the measurements

These different methods have been put into application, mainly thanks to the HST, to construct a reliable astronomical distance scale. In particular, this distance scale has been calibrated up to a level of 5% by cepheids. The results concerning the Hubble diagram are presented in Fig. 4.2. The different values of the Hubble constant obtained by these various methods [7] are summarized in Table 4.1.

Table 4.1 Value of the Hubble constant as measured by the various methods discussed in the text.

| H_0 (km · s ⁻¹ / Mpc) | Method |
|------------------------------------|--------------------|
| $71 \pm 2 \pm 6$ | SN Ia |
| $71 \pm 3 \pm 7$ | Tully–Fisher |
| $70 \pm 5 \pm 6$ | surface brightness |
| $72 \pm 9 \pm 7$ | SN II |
| $82 \pm 2 \pm 6$ | fundamental plane |
| 72 ± 8 | combined methods |

4.1.1.3 Physical methods

In parallel with these astronomical methods, there exist at least three other methods to determine the Hubble constant, avoiding the need to construct a scale of distances.

1. *Expansion of the photosphere of SN II:* the front of the SN II propagates at a velocity of order $v/c \sim 0.01$ after the explosion. This velocity can be measured by the Doppler shift in the supernova spectrum. The radius of the photosphere is then $r_{\text{photosphere}} = v(t - t_{\text{explosion}})$. If the angular diameter θ under which the supernova is observed can be measured, then its distance can be deduced, $D = v(t - t_{\text{explosion}})/\theta$. Unfortunately, the angular diameter is difficult to measure directly for extragalactic supernovæ.
2. *Lensing effect:* the arrival times of two gravitationally lensed images of the same point-like source depend on the path followed by light and on the gravitational potentials along these paths (Chapter 7). The time delay between two light signals together with the angular separation between the two images of a variable quasar allow the Hubble constant to be deduced. However, this method has some problems. The gravitational lens is usually a galaxy, the mass distribution of which is not known independently. Thus, there is a degeneracy between the distribution of the lens mass and the Hubble constant. Ideally, we should need a measurement of the velocity dispersion as a function of the position. Only a dozen known systems, having both a favourable geometry and at least one variable source, is currently known.
3. *Sunyaev–Zel'dovich effect:* the inverse Compton scattering of a microwave background photon (see Section 4.4) by the ionized gas of a cluster, induces a distortion

of the blackbody spectrum, known as the (thermal) Sunyaev–Zel'dovich effect [8]. The scattering probability P can roughly be expressed as a function of the cluster diameter D_{cluster} and its average electronic density n_e by $P \sim n_e \sigma_T D_{\text{cluster}}$, σ_T being the Thomson scattering cross-section.

The temperature and density of the hot gas can be obtained from an X emission map. Using the fact that the X flux, unlike the spectral distortion of the microwave background, depends on the distance to the cluster, the latter can thus be deduced. The main uncertainties of this method come from the heterogeneities of the gas distribution (which tend to lower the measured value of H_0), from the projection effects, and the hydrostatic equilibrium hypothesis.

To conclude, it is reassuring to notice that methods based on very different physical principles lead to consistent values of the Hubble constant (see table 4.2), themselves in agreement with those obtained by astronomical methods.

Table 4.2 Measurement of the Hubble constant by physical methods.

| H_0 (km · s ⁻¹ /Mpc) | Method | Reference |
|-----------------------------------|---------------------------|--------------------------------|
| 73 ± 15 | photosphere | Schmidt <i>et al.</i> (1994) |
| $72 \pm 7 \pm 15$ | lensing effects | Tonry and Franx (1998) |
| 60 ± 20 | lensing effects | Fassnacht <i>et al.</i> (2000) |
| 60 ± 4 | Sunyaev–Zel'dovich effect | Reese <i>et al.</i> (2002) |
| 72 ± 5 | WMAP | Spergel <i>et al.</i> (2003) |

4.1.2 The contribution of supernovæ

As illustrated in section 4.1.1, SN Ia represent to date the most competitive candidates to extend the Hubble diagram to large redshifts, for which the luminosity distance-redshift relation is no longer linear. This offers the opportunity to measure cosmological parameters other than the Hubble constant.

4.1.2.1 *Distant supernovæ*

There exist two families of supernovæ. Those for which the optical spectrum includes hydrogen traces are called type II, whereas those lacking hydrogen are of type I. Moreover, type I supernovæ are subdivided into three very different sub-families (a,b,c). SN Ib and Ic, just as SN II, are created by the explosion of massive stars (called Wolfe–Rayet) and lead to a residual black hole or neutron star. The mass lost by the progenitor is more significant for the Ic than for the Ib so that the latter still have a layer of helium. SN Ia, on the other hand, appear in binary systems where one of the two stars is a white dwarf that accretes the mass of its companion. When it reaches the critical Chandrasekhar mass, $1.4 M_\odot$, the white dwarf collapses and explodes in a supernova. It has been established, thanks to the study of close supernovæ, that their luminosity curve could be calibrated using an empirical relation between the peak and the width of the luminosity curve. Thus, SN Ia seem to be good standard candles that can be observed up to large redshifts.

Two teams have worked in parallel on the search for distant supernovæ: the Supernova Cosmology Project (SCP) [9] and the High-Z Supernova Search Team (HZT) [10]. In 1998, they independently published the Hubble diagram of type Ia supernovæ (see Fig. 4.3). The SCP group used a catalogue of 42 SN Ia composed of 18 SN Ia with $z < 0.101$ and 24 SN Ia with $0.180 < z < 0.830$. The analysis of the HZT group relies on 34 close SN Ia and 16 SN Ia with $0.16 < z < 0.62$. Figure 4.3 summarizes the observations of both teams. Since then, the number and the maximal redshift of the observed supernovæ have increased greatly.

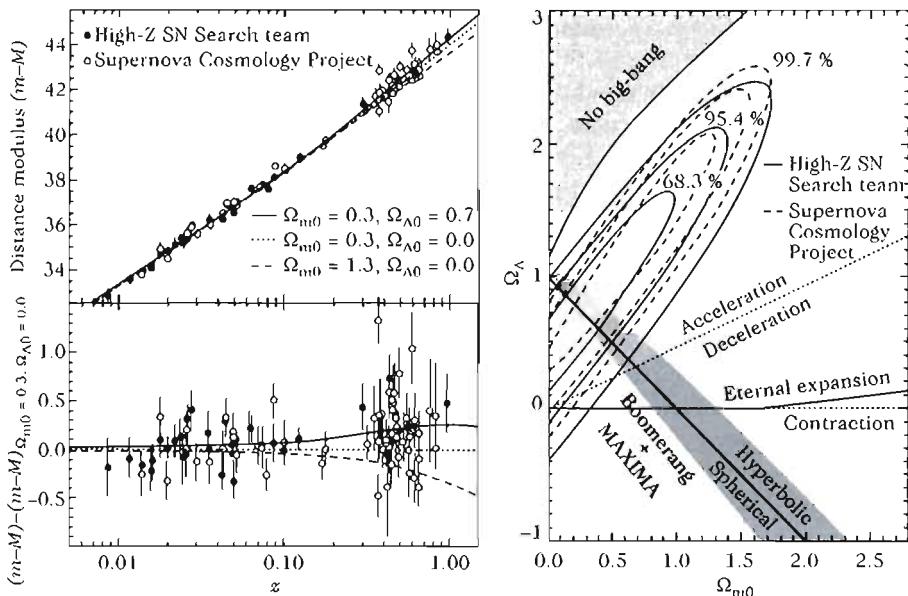


Fig. 4.3 The Hubble diagram for the supernovæ of both HZT and SCP teams compared with the predictions of three Λ CDM models. The bottom panel shows the difference between the data and a Universe with $(\Omega_{m0}, \Omega_{\Lambda0}) = (0.3, 0)$. (right): The constraints on the cosmological parameters $(\Omega_{m0}, \Omega_{\Lambda0})$ obtained by both experiments, compared with those obtained from the cosmological background (see Chapter 6) based on the observations of MAXIMA and BOOMERanG balloons. Adapted from Ref. [11].

4.1.2.2 Implication for the cosmological parameters

As soon as the redshift is not too small, the relation between the distance luminosity and the redshift becomes non-linear and the next-to-leading order term is proportional to $\frac{1}{2}(1-q_0)z^2$ [see (3.99)]. The analysis of both teams proves that the deceleration parameter satisfies

$$q_0 < 0 \quad (4.1)$$

at 2.8σ with no other hypothesis than $\Omega_{m0} > 0$. This conclusion follows from the parameterization (3.92) that does not make any hypothesis on the material composition

of the Universe and relies only on the hypothesis that the Universe can be described by a Friedmann–Lemaître space-time. We can thus consider the fact that the Universe is accelerating to be a robust conclusion of the analysis of SN Ia data.¹

In a Λ CDM Universe, this relation mainly gives information on the two cosmological parameters Ω_{m0} and $\Omega_{\Lambda0}$, as indicated by (3.94), and the constraint (4.1) implies that $\Omega_{\Lambda0} > \frac{1}{2}\Omega_{m0}$. Figure 4.3 summarizes the supernovæ constraints. These data imply that the cosmological constant must be nonvanishing and positive. The ellipses of constraints of Fig. 4.3 can be roughly summarized as

$$8\Omega_{m0} - 6\Omega_{\Lambda0} \simeq -2 \pm 1. \quad (4.2)$$

For tighter constraints, one should combine the data of other observations such as those from the cosmological microwave background. The latter indicates that the Universe is almost flat, $\Omega_{m0} + \Omega_{\Lambda0} \simeq 1.02 \pm 0.02$ according to the analysis of WMAP (see Chapter 6). Thus, for a Λ CDM Universe,

$$\Omega_{\Lambda0} \sim 0.7, \quad \Omega_{m0} \sim 0.3. \quad (4.3)$$

The order of magnitude of this analysis is confirmed by precise statistical analysis.

4.1.2.3 Robustness of the conclusion

The greatest uncertainty lies in the hypothesis that distant supernovæ are standard candles and can be calibrated in the same way as close supernovæ. Three effects should be quantified with precision in order to know if this effect is real or if it is an artefact.

- The behaviour of supernovæ could vary with cosmic time, mainly due to the change of metallicity, i.e. of their chemical composition beyond helium-4. If distant supernovæ have a luminosity peak systematically weaker than closer ones, the conclusion $\Omega_{\Lambda0} > 0$ would be softened. If the explosions of SN Ia are stronger, then the conclusion would be reinforced. It is difficult to obtain information on the intrinsic luminosity and one can more easily study other parameters of the light curve. If two families (close and distant) are in all aspects identical, it is, however, highly probable that their luminosity is also the same.
- The absorption by the interstellar medium could give the illusion that distant supernovæ are systematically dimmer. The analysis of the light curve in several colours can give an estimate of this dimming. The effects of gravitational lensing could also soften the luminosity.
- The possibility of an observational bias should also be considered.

The increasing number of observed supernovæ provides a better understanding of their physics and current studies tend to prove that the acceleration of the Universe can not be explained by any one of these effects [11]. Figure 4.4 compares the astrophysical and cosmological effects.

¹We emphasize that this conclusion can now be reached by the combination of other data (e.g. CMB and weak lensing) without taking the SN Ia into account.

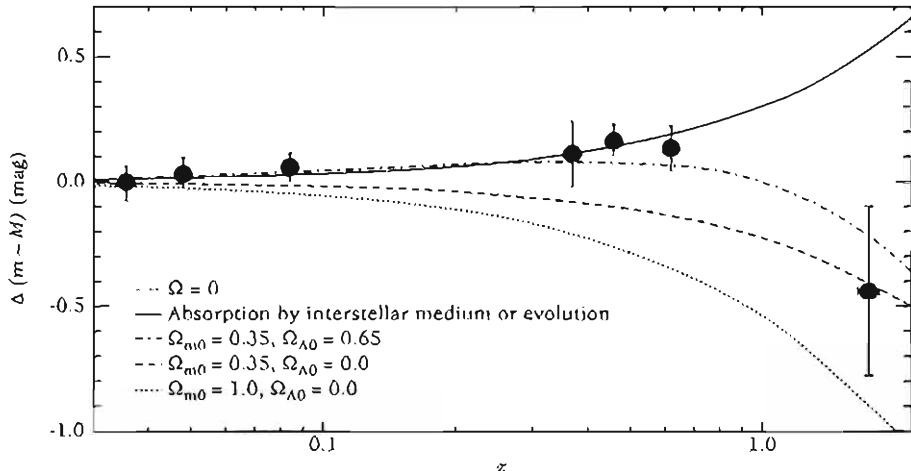


Fig. 4.4 Hubble diagram for type Ia supernovæ compared to the one for an empty Universe $(\Omega_{m0}, \Omega_{\Lambda0}) = (0, 0)$. The measurement of the SN Ia at $z = 1.7$ is incompatible with some astrophysical effects that could mimick an acceleration of the Universe at lower redshifts. Adapted from Ref. [11].

4.1.2.4 Cosmological constant or dark energy

The previous results assume that only a cosmological constant and ordinary matter dominate. However, any kind of matter with equation of state $w < -1/3$ can lead to $\eta_0 < 0$ [see (3.94)].

It is legitimate to wonder whether it can be checked from the data that the matter responsible for the acceleration of the Universe is indeed a cosmological constant, i.e. a component with an equation of state $w = -1$. Figure 4.5 presents the constraints on the pair of parameters (w, Ω_{m0}) for a flat Universe if w is assumed to be constant. For $\Omega_{m0} \sim 0.3$ we obtain

$$w < -0.6. \quad (4.4)$$

The constraints on the equation of state depend on which data have been combined and on the parameterization of the equation of state. Figure 4.5 compares the constraints obtained by combining the observations from the microwave background (WMAP), galaxy catalogues (SDSS), the supernovæ, and from the Lyman- α forest in various ways.

It is currently conventional to call *dark energy* the component responsible for the acceleration of the Universe. This dark energy could be a cosmological constant but other candidates also exist (see Chapter 12 for a more detailed discussion). The majority of these candidates have a dynamical equation of state. The constraints on the equation of state of the dark energy change very rapidly with the flow of new observations. Let us cite, for instance, a constraint [12] on a parameterization of the form $w = w_0 + w_a z/(1+z)$; it leads to

$$w_0 = -0.981^{+0.193}_{-0.193}, \quad w_a = 0.05^{+0.83}_{-0.65}, \quad (4.5)$$

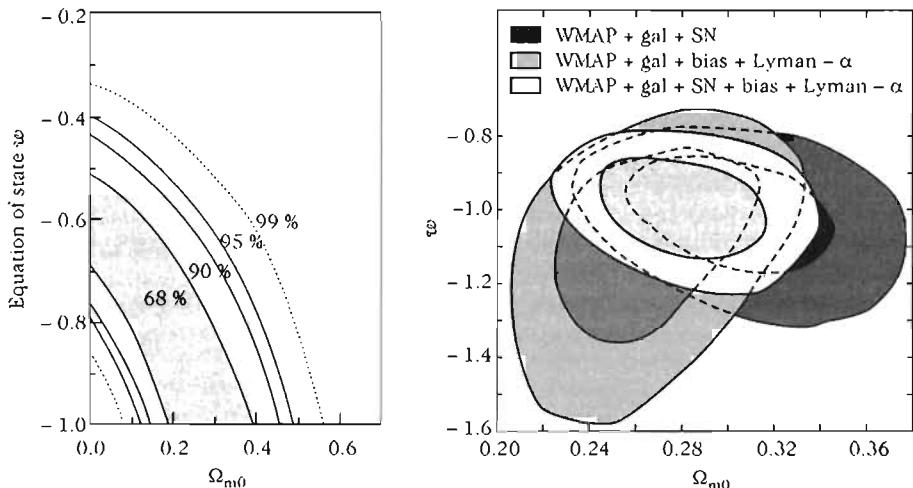


Fig. 4.5 (left): Constraints on a constant equation of state of dark energy as a function of Ω_{m0} for a flat Universe. Adapted from Ref. [9]. (right): Same thing by combining various sets of observation (contours at 68% and 95%), cosmological microwave background (WMAP), galaxies (SDSS), supernovæ and Lyman- α forest, in different ways. Adapted from Ref. [12].

at 1σ . The determination of the nature of this dark energy is one of the important challenges of modern cosmology and will be discussed in the final chapters of this book.

4.1.3 The age of the Universe

A constraint on the cosmological parameters can also be obtained from the measurements of the age of the Universe. A lower bound on this age can then be set that should be compatible with the dynamical age of the Universe computed from the Friedmann equations. Data from supernovæ imply that the dynamical age of the Universe for a flat Λ CDM model is

$$t_0 = 14.9_{-1.1}^{+1.4} \left(\frac{h}{0.63} \right)^{-1} \times 10^9 \text{ years}, \quad t_0 = 14.2_{-0.8}^{+1.0} \left(\frac{h}{0.63} \right)^{-1} \times 10^9 \text{ years}, \quad (4.6)$$

using data, respectively, from SCP [9] and HZT [10].

4.1.3.1 Ages in the Milky Way

The oldest objects in the Milky Way are low metallicity² stars and are located in the halo of the Milky Way. There are three methods to determine the age of these stars [13].

1. *Nucleochronology*: conceptually, this is the simplest method, analogous to radioactive dating methods, and it represents the most direct way to date the Universe,

²By *metallicity*, we mean here the content in chemical elements beyond helium-4. The ratio [Fe/H] traces this metallicity.

t_0 . The age of the stars is estimated from the abundance of long-lived radioactive nuclei for which a half-life is known (for details, see Ref. [14]). The radioisotope ^{232}Th (thorium) with half-life of 14 million years has been used to date the age of the stars of the Galaxy (in particular the star CS 22892) but its abundance is only halved on a period equal to the age of the Universe. In principle, ^{238}U is a better indicator since its half-life is 4.5 million years. The abundance of this isotope was derived by observing the star CS31082-0018 [15], $\log(\text{U}/\text{H}) = -13.7 \pm 0.14$, which corresponds to an age of $(12.5 \pm 3) \times 10^9$ years.

2. *Cooling of white dwarfs*: white dwarfs are the terminal stage of the evolution of stars with mass smaller than around $8M_\odot$. As they get old, white dwarfs get cooler and dimmer. Using a model of their cooling, their cooling rate, and thus the age of the Universe, can be calculated. Unfortunately, since they are not very bright, white dwarfs are very difficult to observe so that the studies are reduced to white dwarfs in the neighbourhood of the Solar System and enable us to estimate the age of the local disk to approximatively $9.8^{+1.1}_{-0.8} \times 10^9$ years.
3. *Main sequence*: theoretical models of stellar evolution make it possible to understand the position of a star in the temperature–luminosity plane. During the main sequence, which corresponds to the longest part of their life, stars burn their hydrogen to produce helium. At the end of this period, the luminosity of the star increases and its surface temperature decreases. It can be shown that the luminosity of the stars at this transition of the main sequence is the quantity least sensitive to errors in the theoretical models and can be used to determine the age of globular clusters up to precision of 7–10%. One should also determine the distance of the cluster that limits the precision of the method (an error of 10% on the distance translates into an error of 20% on the age!). Thanks to the satellite Hipparcos a better calibration of the local distance scale was possible. The age of globular clusters is roughly in the range $(11 – 14) \times 10^9$ years.

Table 4.3 Summary of the main constraints on the age of the Universe.

| Age (10^9 years) | Method | Reference |
|---------------------|--|-------------------------------|
| 15.2 ± 3.7 | nucleochronology (star CS 22892) | Sneden <i>et al.</i> (1996) |
| 12.5 ± 3 | nucleochronology (star CS 31082-0018) | Cayrel <i>et al.</i> (2001) |
| 11.5 ± 1.3 | 5 methods (globular clusters) | Chaboyer <i>et al.</i> (1998) |
| 11.8 ± 1.2 | main sequence (globular clusters) | Gratton <i>et al.</i> (1997) |
| 14 ± 1.2 | main sequence (globular clusters + binary) | Pont <i>et al.</i> (1998) |
| 13.7 ± 0.2 | WMAP + large-scale structure | Spergel <i>et al.</i> (2003) |

4.1.3.2 Age of the Universe

To conclude, the minimum age of our Universe can be estimated by specifying the ages of the oldest objects in our Galaxy. To obtain the age of the Universe, the time taken to form these objects should also be estimated. Unfortunately there is no robust theory

for the beginning of stellar-formation processes. The beginning of star formation is estimated at $z \sim 5 - 20$, which corresponds to a period of $(0.1 - 2) \times 10^9$ years. This allows us to conclude (see Table 4.3) that

$$9.6 \leq \frac{t_u}{10^9 \text{ years}} \leq 15.4. \quad (4.7)$$

4.1.3.3 Compatibility between the Hubble constant and the age of the Universe

Having measured the Hubble constant, the dynamical age of the Universe can be obtained from (3.66) which only depends on the cosmological parameters Ω_{m0} and $\Omega_{\Lambda 0}$ (the duration of the radiation era being negligible).

To understand if the constraints on the Hubble constant and on the measurements of the age of the Universe are compatible, we first present a comparison between data from supernovæ and the theoretical computation of the dynamical age of the Universe (Fig. 4.6). We then compare $H_0 t_u$ with $H_0 t_0(\Omega_{m0}, \Omega_{\Lambda 0})$, assuming that H_0 and t_u are known up to 10%. These measurements are consistent if the cosmological constant does not vanish, as independently indicated by the results from supernovæ. An Einstein-de Sitter Universe is marginally excluded (up to 1.5σ) by these measurements.

4.2 Thermodynamics in an expanding Universe

The previous sections have convinced us that the Universe is expanding. Since radiation scales as a^{-4} while pressureless matter scales as a^{-3} , the Universe was dominated by radiation in the past for redshifts larger than

$$z_{\text{eq}} \simeq 3612 \Theta_{2.7}^{-4} \left(\frac{\Omega_{m0} h^2}{0.15} \right), \quad (4.8)$$

obtained by equating the matter and radiation energy densities and where $\Theta_{2.7} \equiv T_{\text{CMB}}/2.725 \text{ K}$ [see (4.137) below]. Since the temperature scales as $(1+z)$, the temperature at which the matter and radiation densities were equal is $T_{\text{eq}} = T_{\text{CMB}}(1+z_{\text{eq}})$ which is of order

$$T_{\text{eq}} \simeq 5.65 \Theta_{2.7}^{-3} \Omega_{m0} h^2 \text{ eV} = 6.56 \times 10^4 \Theta_{2.7}^{-3} \Omega_{m0} h^2 \text{ K}. \quad (4.9)$$

Above this energy, the matter content of the Universe is in a very different form from that of today. In particular, when the temperature T becomes larger than twice the rest mass m of a charged particle, the energy of a photon is large enough to produce particle-antiparticle pairs. Thus, when $T \gg m_e$, both electrons and positrons were present in the Universe, so that the particle content of the Universe changes during its evolution, while it cools down.

Here, we describe this phase and first focus on the particle distribution in equilibrium and then on some out-of-equilibrium processes leading, for instance, to the production of thermal relics.

4.2.1 Equilibrium thermodynamics

For a description of thermodynamics in the relativistic context and in cosmology, one can refer to Refs. [16–18].

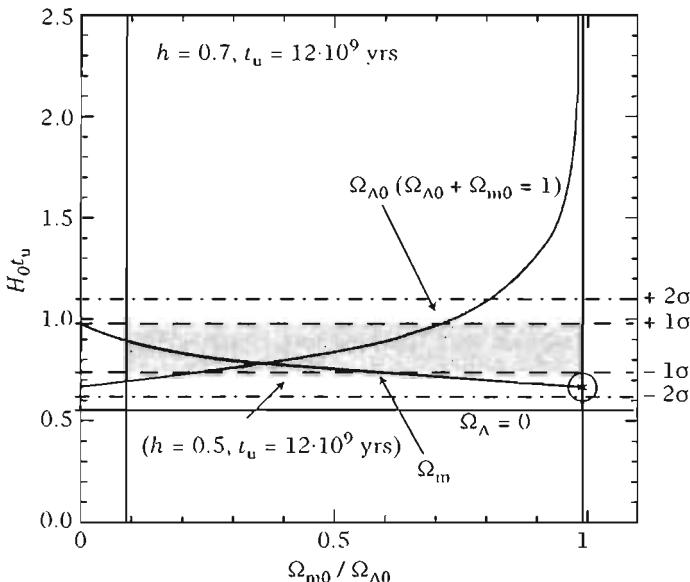


Fig. 4.6 Comparison between the constraints on the product of the Hubble constant and the age of the Universe, $H_0 t_u$, and the product of the Hubble constant and the dynamical age of the Universe, $H_0 t_0$. The latter depends on the cosmological parameters. We have fixed $h = 0.7$ and $t_u = 12 \cdot 10^9$ years. The dark curve represents a flat Universe ($\Omega_{m0} + \Omega_{\Lambda0} = 1$) in which case the abscissa is $\Omega_{\Lambda0}$. The light curve is for a Universe with no cosmological constant ($\Omega_{\Lambda0} = 0$), for which the abscissa is then Ω_{m0} . The grey regions indicate the areas at 1 and 2σ for $H_0 t_u$ and the plain horizontal line is the case where $h = 0.5$. The circle corresponds to an Einstein-de Sitter Universe for which $H_0 t_u = 2/3$. From Ref. [6].

4.2.1.1 Distribution functions and thermodynamical quantities

To describe physical processes during the radiation-dominated era, the distribution function $f_i(p, T)$ of the particles present in the Universe should be determined.

Distribution functions

Particle interactions are mainly characterized by a reaction rate Γ . If this reaction rate is much larger than the Hubble expansion rate, then it can maintain these particles in thermodynamic equilibrium at a temperature T . Particles can thus be treated as perfect Fermi-Dirac and Bose-Einstein gases with distribution³

$$F_i(E, T) = \frac{g_i}{(2\pi)^3} \frac{1}{\exp[(E - \mu_i)/T_i(t)] \pm 1} \equiv \frac{g_i}{(2\pi)^3} f_i(E, T), \quad (4.10)$$

³The distribution function depends a priori on (x, t) and (p, E) but the homogeneity hypothesis implies that it does not depend on x and isotropy implies that it is a function of $p^2 = p^2$. Thus, it follows from the cosmological principle that $f(x, t, p, E) = f(E, t) = f[E, T(t)]$.

where g_i is the degeneracy factor, μ_i is the chemical potential and $E^2 = p^2 + m^2$. The normalization of f_i is such that $f_i = 1$ for the maximum phase-space density allowed by the Pauli principle for a fermion. T_i is the temperature associated with the given species and, by symmetry, it is a function of t alone, $T_i(t)$. Interacting species have the same temperature. Among these particles, the Universe contains an electrodynamic radiation with blackbody spectrum [see Section 4.4] at a temperature of 2.725 K today. Any species interacting with photons will hence have the same temperature as these photons as long as $\Gamma_i \gg H$. The photon temperature $T_\gamma = T$ will thus be called the temperature of the Universe.

If the cross-section behaves as $\sigma \sim E^n \sim T^n$ (for instance, $n = 2$ for electroweak interactions) then the reaction rate behaves as $\Gamma \sim n\sigma \sim T^{n+3}$; the Hubble constant behaves as $H \sim T^2$ in the radiation period. Thus, if $n + 1 > 0$, there will always be a temperature below which the interaction decouples while the Universe cools down. The interaction is then no longer efficient; it is said to be frozen, and can no longer keep the equilibrium of the given species with the other components. This property is at the origin of the thermal history of the Universe and of the existence of relics. This mechanism, during which an interaction can no longer maintain the equilibrium between various particles because of the cosmic expansion, is called *decoupling*. A particle will be said to be coupled if $\Gamma \gtrsim H$ and decoupled if $\Gamma \lesssim H$. This criteria of comparing the reaction rate and the rate H is simple; it often gives a correct order of magnitude, but a more detailed description of the decoupling should be based on a microscopic study of the evolution of the distribution function. As a concrete example, consider Compton scattering. Its reaction rate, $\Gamma_{\text{Compton}} = n_e \sigma_T c$, is of order $\Gamma_{\text{Compton}} \sim 1.4 \times 10^{-3} H_0$ today, which means that, statistically, only one photon in 700 interacts with an electron in a Hubble time today. However, at a redshift $z \sim 10^3$, the electron density is 10^3 times larger and the Hubble expansion rate is of order $H \sim H_0 \sqrt{\Omega_{m0}(1+z)^3} \sim 2 \times 10^4 H_0$ so that $\Gamma_{\text{Compton}} \sim 80H$. This means that statistically at a redshift $z \sim 10^3$ a photon interacts with an electron about 80 times in a Hubble time. This illustrates that backward in time densities and temperature increase and interactions become more and more important.

Note that if the thermal equilibrium can be maintained by interactions within a cosmic plasma, by causality, these interactions cannot establish such an equilibrium between two causally disconnected regions that would have different initial temperatures. The hypothesis of the homogeneity of space implies that the Universe is assumed to be initially in thermal equilibrium. The justification of this state will be discussed in Chapter 8.

Thermodynamical quantities

The distribution function (4.10) can be used to define macroscopic quantities such as the particle number density, n , energy density, ρ and pressure, P , for each species ' i ', as⁴

⁴Since $E^2 - m^2 = p^2$ implies that $pdp = EdE$ and, because of isotropy, $d^3p = 4\pi p^2 dp = 4\pi\sqrt{E^2 - m^2} EdE$.

$$n_i(t) = \int F_i(p, T) d^3 p = \frac{g_i}{2\pi^2} \int_m^\infty \frac{\sqrt{E^2 - m^2} E dE}{\exp[(E - \mu_i)/T] \pm 1}, \quad (4.11)$$

$$\rho_i(t) = \int F_i(p, T) E(p) d^3 p = \frac{g_i}{2\pi^2} \int_m^\infty \frac{\sqrt{E^2 - m^2} E^2 dE}{\exp[(E - \mu_i)/T] \pm 1}, \quad (4.12)$$

$$P_i(t) = \int F_i(p, T) \frac{p^2}{3E} d^3 p = \frac{g_i}{6\pi^2} \int_m^\infty \frac{(E^2 - m^2)^{3/2} dE}{\exp[(E - \mu_i)/T] \pm 1}. \quad (4.13)$$

These three quantities can be expressed in terms of the integrals

$$I_i^{(m,n)}(x_i, y_i) \equiv \int_{x_i}^\infty \frac{u^m (u^2 - x_i^2)^{n/2} du}{\exp[(u - y_i)] \pm 1}, \quad (4.14)$$

with

$$x_i \equiv \frac{m_i}{T}, \quad y_i \equiv \frac{\mu_i}{T}, \quad (4.15)$$

as

$$n_i = \frac{g_i}{2\pi^2} T^3 I_i^{(1,1)}, \quad \rho_i = \frac{g_i}{2\pi^2} T^4 I_i^{(2,1)}, \quad P_i = \frac{g_i}{6\pi^2} T^4 I_i^{(0,3)}. \quad (4.16)$$

Table 4.4 summarizes the expression of these quantities within various limits for fermions (F) and bosons (B), where $\zeta(3) \approx 1.202$ is the value of the Riemann zeta function at 3.

Table 4.4 Summary of the main limits of the thermodynamical quantities (4.11)–(4.13) for fermions (F) and bosons (B).

| Limit | particles | n | ρ | P |
|---------------------|-----------|--------------------------------|----------------------------|----------|
| $T \gg m, \mu$ | B | $g(\zeta(3)/\pi^2)T^3$ | $(\pi^2/30)gT^4$ | $\rho/3$ |
| | F | $g(3\zeta(3)/4\pi^2)T^3$ | $(7\pi^2/8 \times 30)gT^4$ | $\rho/3$ |
| $\mu \gg T \gg m$ | F | $g\mu^3/6\pi^2$ | $g\mu^3/8\pi^2$ | $\rho/3$ |
| $T \gg m, \mu < -T$ | B,F | $(g/\pi^2)e^{\mu/T}T^3$ | $(3g/\pi^2)e^{\mu/T}T^4$ | $\rho/3$ |
| $T \ll m$ | B,F | $g(mT/2\pi)^3/2 e^{(\mu-m)/T}$ | $(m + 3T/2)n$ | nT |

4.2.1.2 Number of relativistic degrees of freedom

The radiation density at a given temperature T can be computed from these expressions (see Table 4.4)

$$\rho_r(T) = g_*(T) \left(\frac{\pi^2}{30} \right) T^4. \quad (4.17)$$

g_* represents the effective number of relativistic degrees of freedom at this temperature,

$$g_*(T) = \sum_{i=\text{bosons}} g_i \left(\frac{T_i}{T} \right)^4 + \frac{7}{8} \sum_{i=\text{fermions}} g_i \left(\frac{T_i}{T} \right)^4. \quad (4.18)$$

The factor 7/8 arises from the difference between the Fermi and Bose distributions. In this regime, the matter content is dominated by radiation and the curvature is negligible, so that the Friedmann equation is of the form

$$H^2 = \frac{8\pi G_N}{3} \left(\frac{\pi^2}{30} \right) g_* T^4. \quad (4.19)$$

Numerically, this amounts to

$$H(T) \cong 1.66 g_*^{1/2} \frac{T^2}{M_p}, \quad (4.20)$$

or equivalently,

$$t(T) \cong 0.3 g_*^{-1/2} \frac{M_p}{T^2} \sim 2.42 g_*^{-1/2} \left(\frac{T}{1 \text{ MeV}} \right)^{-2} \text{ s}, \quad (4.21)$$

obtained by integrating $H = -\dot{T}/T$.

4.2.1.3 Chemical potential and particle–antiparticle asymmetry

Inelastic scattering processes such as Bremsstrahlung $e + p \longleftrightarrow e + p + \gamma$ imply that the number of photons is not conserved and thus their chemical potential should vanish

$$\mu_\gamma = 0. \quad (4.22)$$

This implies that photons must have a Planck distribution.

Any particle, A , kept in chemical equilibrium with its antiparticle by a reaction of the form $A + \bar{A} \longleftrightarrow \gamma + \gamma$ must then satisfy the constraint

$$\mu_A + \mu_{\bar{A}} = 0. \quad (4.23)$$

As long as $T \gg m_A$, using (4.11), this implies an asymmetry

$$n_A - n_{\bar{A}} \simeq \frac{g_A T^3}{6\pi^2} \left[\pi^2 \left(\frac{\mu_A}{T} \right) + \left(\frac{\mu_A}{T} \right)^3 \right] \quad (4.24)$$

between particles and antiparticles. At lower temperatures ($T \ll m_A$) this asymmetry is exponentially suppressed

$$n_A - n_{\bar{A}} \simeq 2g_A \left(\frac{m_A T}{2\pi} \right)^{3/2} e^{-m_A/T} \sinh \left(\frac{\mu_A}{T} \right). \quad (4.25)$$

When T becomes smaller than m_A , A and \bar{A} annihilate and only this small excess survives. This is, for instance, the case for the electrons.

Note that the electrical neutrality of the Universe implies that the number of protons n_p is equal to $n_e - n_{\bar{e}}$. Using the constraint $n_p/n_\gamma \sim 5 \times 10^{-10}$ [see Section 4.3, (4.142)], we obtain

$$\frac{n_p}{n_\gamma} \sim \frac{g_e}{g_\gamma} \frac{\pi^2}{6\zeta(3)} \frac{\mu_e}{T} \sim 10^{-8}, \quad (4.26)$$

so that $1.34\mu_e/T \sim 6.5 \times 10^{-10}$. The chemical potential of the electrons and positrons can thus be neglected.

4.2.1.4 Entropy

In order to follow the evolution of the processes, it is convenient to have conserved quantities such as the entropy.

Combining the conservation equation (3.27), rewritten in the form $d(\rho a^3) = -Pda^3$, and the derivative of the pressure (4.13) with respect to the temperature,

$$\frac{dP}{dT} = \frac{\rho + P}{T} + nT \frac{d}{dT} \left(\frac{\mu}{T} \right),$$

we get that the quantity⁵

$$s \equiv \frac{\rho + P - n\mu}{T} \quad (4.27)$$

satisfies⁶ the equation $d(sa^3) = -(\mu/T)d(na^3)$. Thus, the product sa^3 is constant (i) as long as matter is neither destroyed nor created, since then na^3 is constant, or (ii) for non-degenerate relativistic matter, $\mu/T \ll 1$. We thus deduce that in the cases relevant for cosmology,

$$d(sa^3) = 0. \quad (4.28)$$

To interpret this quantity, let us consider the expression (4.27) which implies that $Td(sa^3) \simeq d(\rho a^3) + Pda^3$, in the limit $\mu/T \ll 1$. We recognise $S = sa^3$ as the entropy and thus s is the entropy density. Hence, (4.28) implies that $s \propto a^{-3}$ as long as $\mu/T \ll 1$. Then, using (4.27), the entropy can be expressed as

$$s = \frac{2\pi^2}{45} q_* T^3, \quad (4.29)$$

where q_* is defined by

$$q_*(T) = \sum_{i=\text{bosons}} g_i \left(\frac{T_i}{T} \right)^3 + \frac{7}{8} \sum_{i=\text{fermions}} g_i \left(\frac{T_i}{T} \right)^3. \quad (4.30)$$

If all relativistic particles are at the same temperature, $T_i = T$, then $q_* = g_*$. Note also that $s = q_* \pi^4 / 45 \zeta(3) n_\gamma \sim 1.8 q_* n_\gamma$, so that the photon number density gives a measure of the entropy.

⁵In the case of a vanishing chemical potential, this expression can be easily derived by starting from the general expression of the entropy of a system as $TdS = dE + PdV$. If we choose T and V as our variables to describe the thermodynamical state of the system then, since they are intensive quantities, P and ρ are functions of T alone; $\rho(T)$ and $P(T)$. It follows that $dE = d(\rho V) = \rho dV + V(d\rho/dT)dT$. Thus,

$$dS = \frac{\rho + P}{T} dV + \frac{V}{T} \frac{d\rho}{dT} dT.$$

We deduce that $(\partial S/\partial V)_T = (\rho + P)/T$ and $(\partial S/\partial T)_V = (V/T)(d\rho/dT)$. The integrability condition then implies that $d\rho/dT = (\rho + P)/T$. Plugging these expressions in dS we deduce that

$$dS = d \left[\frac{V}{T} (\rho + P) \right],$$

from which we get the expression of the entropy up to an additive constant.

⁶Starting from $TdS = dE + PdV - \mu dN$ and using that $V \propto a^3$, so that $E \propto \rho a^3$, and $S \propto sa^3$ we deduce that $Td(sa^3) = d(\rho a^3) + Pda^3 - \mu d(na^3)$. Taking into account that the first two terms of the r.h.s. cancel, we deduce the evolution equation of sa^3 .

The number of particles in a given volume, $N \propto na^3$, is constant if no matter is created or annihilated. The relation (4.28) thus implies that

$$Y = \frac{n}{s} \quad (4.31)$$

represents the number of particles per comoving volume. For relativistic particles, this quantity remains constant during the evolution. It can be easily computed in two limiting cases

$$Y = \frac{45}{2\pi^4} \zeta(3) \frac{g_*}{q_*} \quad \text{if } T \gg \mu, m, \quad (4.32)$$

$$= \frac{45}{4\sqrt{2}\pi^5} \frac{g_*}{q_*} \left(\frac{m}{T}\right)^{3/2} \exp\left[-\frac{(m-\mu)}{T}\right] \quad \text{if } T \ll m. \quad (4.33)$$

4.2.1.5 Decoupled species

If at a given time, the interaction rate is no longer high enough to maintain the thermodynamic equilibrium of the species i , i.e. if $\Gamma_i \lesssim H$, then this species evolves independently from the others. If this species was in thermodynamic equilibrium before decoupling, then its distribution function at the time of decoupling, t_D , is

$$f_i(p, t_D) = \frac{1}{\exp[(E - \mu_i)/T_i(t_D)] \pm 1}. \quad (4.34)$$

After decoupling, particles propagate freely and the form of the distribution function is conserved. Only the momentum is redshifted by the expansion as $p(t) = p(t_D)a_D/a(t)$. It follows that the distribution function at any $t > t_D$ is given by

$$f_i(p, t) = f_i\left[\frac{a(t)}{a_D} p, t_D\right]. \quad (4.35)$$

Thus, if a species was in thermodynamic equilibrium at a given time in the history of the Universe, one can determine its distribution at any time.

If the decoupling occurred when the particle was relativistic, ($T \gg m, \mu$), then the distribution function (4.35) takes the simple form

$$f_i(p, t > t_D) = \frac{1}{\exp[E/T_i(t)] \pm 1}, \quad T_i(t) = T_D \frac{a_D}{a(t)}, \quad (4.36)$$

where $T_D = T(t_D)$. The temperature of this decoupled species always decreases as a^{-1} and its entropy S_i is conserved separately. If this particle becomes non-relativistic after $t_{NR} \gg t_D$, we have $E(t) \simeq m$ for any $t > t_{NR}$ but the distribution function keeps the form (4.36).

If the decoupling occurred when the particle was non-relativistic ($T \ll m$), then $E \simeq m + p^2/2m$ and the distribution function is given by

$$f_i(p, t > t_D) = e^{-(m-\mu)/T_D} e^{-p^2/2mT_i(t)}, \quad T_i(t) = T_D \left[\frac{a_D}{a(t)}\right]^2. \quad (4.37)$$

Thus, everything goes as if the particles had an effective chemical potential, $\mu(t) = m + [\mu_D - m]T(t)/T_D$. We then obtain

$$\rho_i = n_i m, \quad n_i \simeq g_i \left(\frac{m T_D}{2\pi} \right)^{3/2} \left(\frac{a_D}{a} \right)^3 e^{-m/T_D}. \quad (4.38)$$

4.2.1.6 Temperatures and their evolution

Temperature of the Universe

Since the total entropy is constant, $q_*(T)T^3a^3$ remains constant during the evolution of the Universe. If, moreover, a species i decouples from the cosmic plasma at temperature T_D then the entropy of this species is conserved on its own so that

$$S = \frac{2\pi^2}{45} q_*(T)T^3a^3 \quad \text{and} \quad S_i = \frac{2\pi^2}{45} q_i(T_i)T_i^3a^3 \quad (4.39)$$

are both constant. The entropy, $S - S_i = (2\pi^2/45)q_\gamma(T)T^3a^3$, of the particles in thermal equilibrium with the photons, is thus also constant. We thus conclude that the temperature of the Universe varies as

$$T \propto q_\gamma^{-1/3}a^{-1}, \quad (4.40)$$

where q_γ is given by the relation (4.30) but only summing over the relativistic particles that are in thermal equilibrium with the photons.

Temperature of a decoupled species

As long as the number of relativistic species does not vary, the temperature decreases as a^{-1} . When a species becomes non-relativistic, its entropy is transferred to the other relativistic particles remaining in thermal equilibrium and therefore see their temperature suddenly increase (q_* is a decreasing quantity) as

$$T(+) = \left[\frac{q_\gamma(-)}{q_\gamma(+)} \right]^{1/3} T(-), \quad (4.41)$$

where $+$ and $-$ refer, respectively, to quantities evaluated after and before the decoupling. At the time of decoupling $T = T_i = T_D$ so that comparing the quantities (4.39) at temperatures T_D and $T < T_D$, we get that the temperature of the decoupled species is related to that of the Universe by

$$\frac{T_i}{T} = \left[\frac{q_i(T_D)}{q_i(T)} \frac{q_\gamma(T)}{q_\gamma(T_D)} \right]^{1/3}. \quad (4.42)$$

This expression generalizes (4.36) to the case where $q_i(T)$ is not constant. In particular, $q_i(T)$ can vary if the species i has a new interaction allowing it to annihilate to produce another particle that does not interact with the cosmic plasma.

4.2.1.7 Neutrinos as an example

Neutrinos are in equilibrium with the cosmic plasma as long as the reactions $\nu + \bar{\nu} \longleftrightarrow e + \bar{e}$ and $\nu + e \longleftrightarrow \nu + e$ can keep them coupled. Since neutrinos are not charged, they do not interact directly with photons.

The cross-section of weak interactions is given by $\sigma \sim G_F^2 E^2 \propto G_F^2 T^2$ as long as the energy of the neutrinos is in the range $m_e \ll E \ll m_w$. The interaction rate is thus of the order of $\Gamma = n\langle\sigma v\rangle \simeq G_F^2 T^5$. We obtain that

$$\Gamma \simeq \left(\frac{T}{1 \text{ MeV}} \right)^3 H, \quad (4.43)$$

where we have used (4.20). Thus, close to $T_D \sim 1 \text{ MeV}$, neutrinos decouple from the cosmic plasma. For $T < T_D$, the neutrino temperature decreases as $T_\nu \propto a^{-1}$ and remains equal to the photon temperature.

Slightly after decoupling, the temperature becomes smaller than m_e . Between T_D and $T = m_e$ there are 4 fermionic states (e^- , e^+ , each having $g_e = 2$) and 2 bosonic states (photons with $g_\gamma = 2$) in thermal equilibrium with the photons. We thus have that

$$q_\gamma(T > m_e) = \frac{7}{8} \times 2 \times 2 + 2 = \frac{11}{2}. \quad (4.44)$$

For $T < m_e$ only the photons contribute to q_γ and hence

$$q_\gamma(T < m_e) = 2. \quad (4.45)$$

The conservation of entropy implies that after $\bar{e} - e$ annihilation, the temperatures of the neutrinos and the photons are related by

$$T_\gamma = \left(\frac{11}{4} \right)^{1/3} T_\nu. \quad (4.46)$$

Thus, the temperature of the Universe is increased by about 40% compared to the neutrino temperature during the annihilation. Since $n_\nu = (3/11)n_\gamma$, there must exist a cosmic background of neutrinos with a density of 112 neutrinos per cubic centimeter and per family, with a temperature of around 1.95 K. The energy density of these neutrinos, as long as they are relativistic, is $\Omega_\nu = (7/8)(4/11)^{4/3}\Omega_\gamma$ per family of neutrino. Using the value (4.141) of $\Omega_{\gamma 0}$, we conclude that today

$$\Omega_{\nu 0} h^2 = 1.68 \times 10^{-5} \left(\frac{N_\nu}{3} \right) \Theta_{2.7}^4, \quad (4.47)$$

if their mass is smaller than 10^{-4} eV.

4.2.1.8 Number of relativistic degrees of freedom

The $\bar{e} - e$ annihilation is the last annihilation to happen, so that knowing T_ν/T_γ gives the values of g_* and q_* today

$$g_{*0} = 2 + 3 \times \frac{7}{8} \times 2 \times \left(\frac{4}{11} \right)^{4/3} = 3.36, \quad q_{*0} = 2 + 3 \times \frac{7}{8} \times 2 \times \left(\frac{4}{11} \right) = 3.91 \quad (4.48)$$

if there are 3 families of massless neutrinos.

The functions $g_*(T)$ and $g_*(T)$ depend on the particle content of our Universe. As an example, Table 4.5 and Fig. 4.7 summarize their evolution in the context of the standard model of particle physics, discussed in Chapter 2.

Table 4.5 Particle content of the Universe as a function of the temperature in the context of the standard model of particle physics. $T_c^{qh} \approx 150\text{--}400$ MeV characterizes the quark–hadron phase transition, which is assumed to be adiabatic. The thermodynamic history of this transition is not yet well understood. This table also assumes that the Higgs boson has a mass larger than the mass of the W^\pm, Z^0 bosons and of the top quark (t). The last three lines of this table are not certain since the Higgs sector and the electroweak transition are not well understood. At larger temperatures, g_* depends greatly on the theoretical model for the fundamental interactions. For instance, for a minimal SU(5) grand unification model, $g_*(T > T_{GUT}) = 647/4$ and in the case of the minimal supersymmetric model, $g_*(T < T_{SUSY}) = 915/4$ since the number of degrees of freedom is almost doubled.

| T | Threshold (GeV) | Relativistic particles | g_* |
|----------------------|--------------------------|---|-------|
| $< m_e$ | 0.511×10^{-3} | γ (+ 3 decoupled ν) | 2 |
| $m_e - T_D(\nu)$ | $(2 - 4) \times 10^{-3}$ | + e^\pm | 11/2 |
| $T_D(\nu) - m_\mu$ | 0.106 | ν start to interact | 43/4 |
| $m_\mu - m_\pi$ | 0.135 | + μ^\pm | 57/4 |
| $m_\pi - T_c^{qh}$ | | + π^\pm, π^0 | 69/4 |
| $T_c^{qh} - m_s$ | 0.194 | $\gamma, 3\nu$'s, e^\pm, μ^\pm $u, \bar{u}, d, \bar{d}, 8 g$'s | 205/4 |
| $m_s - m_c$ | 1.27 ± 0.05 | + s, \bar{s} | 247/4 |
| $m_c - m_\tau$ | 1.78 | + c, \bar{c} | 289/4 |
| $m_\tau - m_b$ | 4.25 ± 0.10 | + τ^\pm | 303/4 |
| $m_b - m_W$ | 80.3 ± 0.3 | + b, \bar{b} | 345/4 |
| $m_W - m_t$ | 180 ± 12 | + W^\pm, Z^0 | 381/4 |
| $m_t - m_{H^0}$ | | + t, \bar{t} | 423/4 |
| $m_{H^0} - T_c^{EW}$ | 300 (?) | + H^0 | 427/4 |

4.2.2 Out-of-equilibrium thermodynamics

The evolution of a decoupled species can thus easily be described. However, the description of the decoupling, or of a freeze-out of an interaction, is a more complex problem that requires us to go beyond the equilibrium description that has been developed previously.

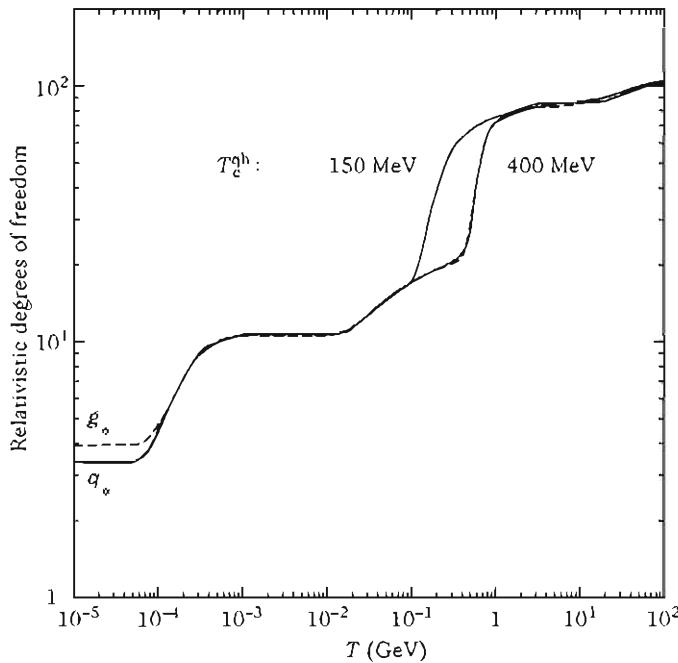


Fig. 4.7 Evolution of the functions g_* and q_* as a function of the temperature in the standard model of particle physics $SU(3)_c \times SU(2)_L \times U(1)_Y$ with two hypotheses concerning the critical temperature of the quark-hadron transition, T_c^{qh} .

4.2.2.1 Boltzmann equation

Evolution of the distribution function

The evolution of the distribution function is obtained from the Boltzmann equation

$$L[f] = C[f], \quad (4.49)$$

where C describes the collisions and $L = d/ds$ is the Liouville operator, with s the length along a worldline. The operator L is a function of eight variables taking the explicit form

$$L[f] = p^\alpha \frac{\partial}{\partial x^\alpha} - \Gamma_{\beta\gamma}^\alpha p^\beta p^\gamma \frac{\partial}{\partial p^\alpha}. \quad (4.50)$$

In a homogeneous and isotropic space-time, f is only a function of the energy and time, $f(E, t)$, so that

$$L[f] = E \frac{\partial f}{\partial t} - H p^2 \frac{\partial f}{\partial E}. \quad (4.51)$$

Using the definition (4.11) of the particle density, and integrating this equation with respect to the momentum p , we obtain⁷

⁷In order to evaluate the second term of (4.51), we use that

$$\dot{n}_i + 3Hn_i = \mathcal{C}_i, \quad \mathcal{C}_i = \frac{g_i}{(2\pi)^3} \int C [f_i(p_i, t)] \frac{d^3 p_i}{E_i}. \quad (4.52)$$

The difficult part lies in the modelling and the evaluation of the collision term. Here, we restrain ourselves to the simple case of an interaction of the form

$$i + j \longleftrightarrow k + l, \quad (4.53)$$

for which the collision term can be decomposed as $\mathcal{C}_i = \mathcal{C}_{kl \rightarrow ij} - \mathcal{C}_{ij \rightarrow kl}$ with

$$\begin{aligned} \mathcal{C}_{ij \rightarrow kl} &= (2\pi)^4 \int \frac{g_i d^4 p_i}{(2\pi)^3} \delta(E_i - p_i^2 - m_i^2) \dots \frac{g_l d^4 p_l}{(2\pi)^3} \delta(E_l - p_l^2 - m_l^2) \\ &\times \delta^{(4)}(p_i + p_j - p_k - p_l) |\mathcal{M}|_{ij \rightarrow kl}^2 f_i f_j (1 \pm f_k) (1 \pm f_l), \end{aligned} \quad (4.54)$$

with a + sign for bosons and a - sign for fermions. $|\mathcal{M}|_{ij \rightarrow kl}$ are the matrix elements describing the interaction. The Dirac delta function imposes the conservation of momentum and of energy. This form also shows that the probability for i to disappear is proportional to $f_i f_j$, i.e. roughly to the density of the interacting species.⁸ The factors $(1 \pm f_k)$ arise from quantum mechanics and are related to the Pauli exclusion principle for fermions and to stimulated emission for bosons.

If \hat{CP} invariance holds, as we assume here, then $\mathcal{C}_{kl \rightarrow ij}$ and $\mathcal{C}_{ij \rightarrow kl}$ involve a unique matrix element, $|\mathcal{M}|^2$, determined by the physical process. Indeed, this invariance implies that the process we consider is reversible and thus that $i + j \rightarrow k + l$ and $k + l \rightarrow i + j$ have the same matrix elements. It follows that

$$\begin{aligned} \mathcal{C}_i &= (2\pi)^4 \int \delta^{(4)}(p_i + p_j - p_k - p_l) \frac{g_i d^3 p_i}{2(2\pi)^3 E_i} \dots \frac{g_l d^3 p_l}{2(2\pi)^3 E_l} \times \\ &\times |\mathcal{M}|^2 [f_k f_l (1 \pm f_i) (1 \pm f_j) - f_i f_j (1 \pm f_k) (1 \pm f_l)]. \end{aligned} \quad (4.55)$$

Equation (4.52) thus involves three sources of evolution for the number density n_i , namely dilution ($3Hn_i$), creation ($\mathcal{C}_i = \mathcal{C}_{kl \rightarrow ij}$) and destruction ($\mathcal{C}_i = \mathcal{C}_{ij \rightarrow kl}$).

Evolution of the particle density

In cosmologically interesting situations, $E - \mu \gg T$. Quantum effects can thus be neglected and $1 \pm f \simeq 1$. Equation (4.52) then takes the form

$$\int \frac{p^2}{E} \frac{\partial f}{\partial E} d^3 p = -3 \int f d^3 p$$

by integrating by parts.

⁸For this reason, we may decompose $\mathcal{C}_{ij \rightarrow kl}$ as $\mathcal{C}_{ij \rightarrow kl} = n_i n_j c_{ij \rightarrow kl}$. Since $c_{ij \rightarrow kl}$ has dimensions of a volume divided by a time, we can interpret it as the average cross-section multiplied by the relative speed of the interacting particles, that is $c_{ij \rightarrow kl} = \langle \sigma_{ij \rightarrow kl} v_{ij} \rangle$. If we define the reaction rate by $\Gamma_{ij \rightarrow kl} = n_j \langle \sigma_{ij \rightarrow kl} v_{ij} \rangle$ then (4.52) takes the form

$$\dot{n}_i + 3Hn_i = n_k \Gamma_{kl \rightarrow ij} - n_i \Gamma_{ij \rightarrow kl}.$$

$$\dot{n}_i + 3Hn_i = \frac{g_i \cdots g_l}{(2\pi)^8} \int \frac{d^3 p_i}{2E_i} \cdots \frac{d^3 p_l}{2E_l} \delta^{(4)}(p_i + p_j - p_k - p_l) |\mathcal{M}|^2 (f_k f_l - f_i f_j). \quad (4.56)$$

In this limit, the distribution functions are of the form $f \propto \exp[(\mu - E)/T]$ so that the particle density (4.11) can be expressed as a function of that at $\mu = 0$ as

$$n_i = e^{\mu_i/T} \bar{n}_i, \quad \bar{n}_i \equiv n_i[\mu_i = 0]. \quad (4.57)$$

Furthermore, the conservation of energy implies that $E_k + E_l = E_i + E_j$ such that the term $f_k f_l - f_i f_j$ takes the form

$$e^{-(E_k + E_l)/T} \left[e^{(\mu_k + \mu_l)/T} - e^{(\mu_i + \mu_j)/T} \right] = e^{-(E_k + E_l)/T} \left(\frac{n_k n_l}{\bar{n}_k \bar{n}_l} - \frac{n_i n_j}{\bar{n}_i \bar{n}_j} \right).$$

The Boltzmann equation (4.56) can thus be written as

$$\dot{n}_i + 3Hn_i = -\langle \sigma v \rangle \left(n_i n_j - \frac{\bar{n}_i \bar{n}_j}{\bar{n}_k \bar{n}_l} n_k n_l \right), \quad (4.58)$$

where $\langle \sigma v \rangle$ is defined as

$$\bar{n}_i \bar{n}_j \langle \sigma v \rangle \equiv \int \frac{d^3 p_i}{2E_i} \cdots \frac{d^3 p_l}{2E_l} \delta^{(4)}(p_i + p_j - p_k - p_l) |\mathcal{M}|^2 \frac{e^{-(E_i + E_j)/T}}{(2\pi)^8}. \quad (4.59)$$

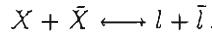
This equation is very general, but to go further we should know the form of the matrix element $|\mathcal{M}|$. We simply note that when $n_i \langle \sigma v \rangle \gg H$, we recover the fact that (4.58) can only be satisfied if the term in brackets cancels, i.e. if

$$\frac{n_i n_j}{\bar{n}_i \bar{n}_j} = \frac{n_k n_l}{\bar{n}_k \bar{n}_l} \iff \mu_i + \mu_j = \mu_k + \mu_l. \quad (4.60)$$

The system is then in chemical equilibrium. This equality is broken as soon as the system is out of equilibrium.

4.2.2.2 Frozen interactions and relic particles

A massive particle X is in thermodynamic equilibrium with its antiparticle \bar{X} for temperatures larger than its mass. Assuming this particle is stable, then its density can only be modified by annihilation or inverse annihilation



If this particle had remained in thermodynamic equilibrium until today, its relic density, $n \propto (m/T)^{3/2} \exp(-m/T)$ would be completely negligible. The relic density of this massive particle, i.e. the residual density once the annihilation is no longer efficient, will actually be more important since, in an expanding space, annihilation cannot keep particles in equilibrium during the whole history of the Universe. That particle is usually called a *relic*.

To evaluate the factor $(f_X f_{\bar{X}} - f_l f_{\bar{l}})$, note that l and \bar{l} are in thermal equilibrium so that, after neglecting their chemical potentials, $f_l = \exp(-E_l/T)$ and $f_{\bar{l}} = \exp(-E_{\bar{l}}/T)$. Conservation of energy imposes $E_X + E_{\bar{X}} = E_l + E_{\bar{l}}$, so that

$$f_l f_{\bar{l}} = \exp\left(-\frac{E_X + E_{\bar{X}}}{T}\right) = \bar{f}_X \bar{f}_{\bar{X}}.$$

The Boltzmann (4.58) now takes the simplified (integrated) form

$$\dot{n}_X + 3Hn_X = -\langle\sigma v\rangle(n_X^2 - \bar{n}_X^2), \quad (4.61)$$

where $\langle\sigma v\rangle$ is defined by (4.59). The conservation of entropy implies that $\dot{n}_X + 3Hn_X = s\dot{Y}_X$, where $Y_X = n_X/s$ is the number of X particles per comoving volume, so that (4.61) can be written as

$$\dot{Y}_X = -\langle\sigma v\rangle s(Y_X^2 - \bar{Y}_X^2). \quad (4.62)$$

Using the variable x [see (4.15) for its definition] and the fact that $dx/dt = Hx$, this equation can be rewritten as

$$\frac{dY_X}{dx} = -\langle\sigma v\rangle \frac{s}{Hx} (Y_X^2 - \bar{Y}_X^2). \quad (4.63)$$

During the radiation era, $H \propto x^{-2}$, so that introducing $\Delta \equiv Y_X - \bar{Y}_X$, and expressing H and s , respectively, by (4.19) and (4.29), it takes the final form

$$\frac{d\Delta}{dx} = -\frac{d\bar{Y}_X}{dx} - \lambda x^{-2} \Delta (\Delta + 2\bar{Y}_X), \quad (4.64)$$

with

$$\lambda = \langle\sigma v\rangle \sqrt{\frac{\pi}{45}} \frac{g_*}{g_*^{1/2}} m M_P. \quad (4.65)$$

To go further, we need the form of $\langle\sigma v\rangle$ and the value of \bar{Y}_X . We illustrate this mechanism in the next section.

4.2.3 Two limiting cases

4.2.3.1 Cold relics

For a cold relic, the decoupling occurs when the particle is non relativistic. Using our previous results to express n_X [(4.11)] and s [(4.29)], we get that

$$\bar{Y}_X = \frac{45}{2\pi^4} \sqrt{\frac{\pi}{8}} \frac{g_X}{g_*} x^{3/2} e^{-x}. \quad (4.66)$$

The quantities $\langle\sigma v\rangle$ and λ depend a priori on the velocity. In general, $\langle\sigma v\rangle \propto v^{2n}$ with $n = 0$ (annihilation of s waves) or $n = 1$ (annihilation of p waves). This quantity can be parameterized in a general way as

$$\langle\sigma v\rangle = \sigma_0 f(x), \quad (4.67)$$

where $f(x)$ is a function of m/T only since $\langle v^2 \rangle \propto T$. Equation (4.64) can be integrated numerically, but the main properties of its solutions can be obtained analytically.

Before decoupling, ($1 < x \ll x_f$), $Y_X \sim \bar{Y}_X$ and Δ and its derivative remain very small. Thus, (4.64) has the approximate solution

$$\Delta \simeq -\frac{1}{2\lambda_0} \frac{x^2}{f(x)} \frac{d \ln \bar{Y}_X}{dx}, \quad \lambda_0 = \sigma_0 \sqrt{\frac{\pi}{45}} \frac{g_*}{g_*^{1/2}} m M_p. \quad (4.68)$$

Since $d \ln \bar{Y}_X / dx = -1 + 3/2x \sim -1$ for $x \gg 1$, we get that before decoupling

$$\Delta \simeq \frac{1}{2\lambda_0} \frac{x^2}{f(x)}. \quad (4.69)$$

After decoupling ($x \gg x_f$), \bar{Y}_X decreases exponentially and becomes negligible in comparison to $Y \sim \Delta$. Equation (4.64) then takes the form

$$\frac{d\Delta}{dx} = -\lambda_0 \frac{\Delta^2}{f(x)x^2}. \quad (4.70)$$

Integrating this equation between x_f and $x = \infty$, we get

$$\Delta_\infty^{-1} - \Delta_f^{-1} = \lambda_0 \int_{x_f}^{\infty} \frac{f(x)}{x^2} dx. \quad (4.71)$$

It follows that

$$Y_\infty \simeq \left[\lambda_0 \int_{x_f}^{\infty} \frac{f(x)}{x^2} dx \right]^{-1}. \quad (4.72)$$

x_f should now be determined. The decoupling happens when Y_X differs significantly from \bar{Y}_X , i.e. when $\Delta(x_f) = c\bar{Y}_X(x_f)$, c being a number of order unity that can be determined by comparison with a numerical integration. By estimating $\Delta(x_f)$ with its value (4.69), x_f is the solution of

$$\frac{x_f^{1/2} e^{x_f}}{f(x_f)} = \frac{45}{\pi^4} \sqrt{\frac{\pi}{8}} \frac{g_X}{g_*^{1/2}} c \lambda_0. \quad (4.73)$$

A similar expression is obtained using the criteria $\Gamma(x_f) = H(x_f)$, which gives

$$\frac{g_X}{g_*^{1/2}} x_f^{1/2} e^{-x_f} f(x_f) \sigma_0 m M_p = 4\pi^3 \sqrt{\frac{2}{45}}. \quad (4.74)$$

From today's entropy density [see (4.140)], we can deduce the relic density of the particle X , $n_{X0} = s_0 Y_\infty$

$$n_{X0} = 1.09 \times 10^4 \frac{g_*^{1/2}}{g_*} \left(\frac{g_{*0}}{3.91} \right) \Theta_{2.7}^3 \frac{\left[\int_{x_f}^{\infty} f(x)/x^2 dx \right]^{-1}}{m M_p \sigma_0} \text{cm}^{-3}. \quad (4.75)$$

To give an estimate of this relic density, let us assume that $f(x) = x^{-n}$ and that $q_*(x_f) \simeq g_*(x_f)$. Since $\rho_{X0} = m n_{X0}$ and $\Omega_X = \rho_X / \rho_{\text{crit}}$, we get

$$\Omega_{X0} h^2 \simeq 0.31 \left[\frac{g_*(x_f)}{100} \right]^{-1/2} (n+1) x_f^{n+1} \left(\frac{q_{*0}}{3.91} \right) \Theta_{2.7}^3 \left(\frac{\sigma_0}{10^{-38} \text{ cm}^2} \right)^{-1}. \quad (4.76)$$

Since x_f depends only weakly on the mass, these expressions give a good estimate of the relic density as a function of the particle's mass and cross-section. Figure 4.8 depicts the numerical integration of the Boltzmann equation (4.64). We recover both asymptotic regimes discussed in this section.

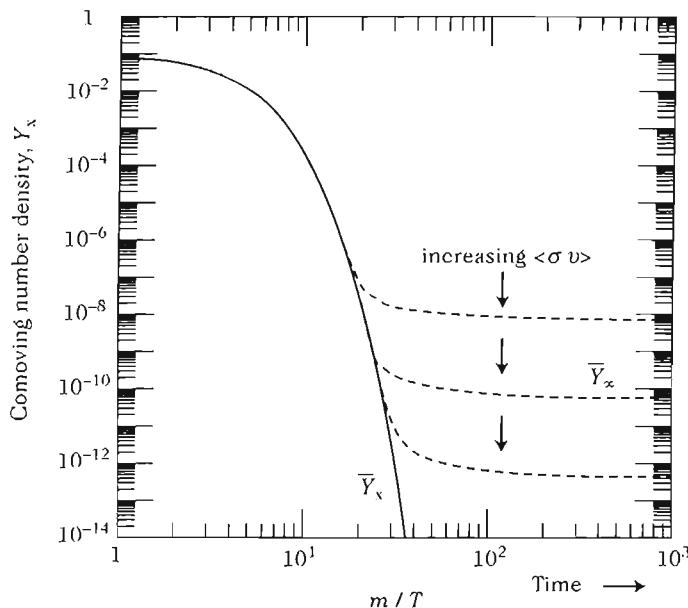


Fig. 4.8 Numerical integration of (4.64). We recover both asymptotic regimes and the dependence in $\langle \sigma v \rangle$ of the relic density Y_X .

4.2.3.2 Hot relics

A hot relic decouples from the plasma while it is relativistic, so that

$$\bar{Y}_X = \frac{45\zeta(3)}{2\pi^4} \varepsilon_{\text{FB}} \frac{g_X}{q_*}, \quad (4.77)$$

where $\varepsilon_{\text{FB}} = 1$ (boson) or $3/4$ (fermion). The interaction is no longer efficient when \bar{Y}_X is constant, so that the relic density is simply given by its value at equilibrium

$$Y_\infty = \bar{Y}_X(x_f) = \frac{45\zeta(3)}{2\pi^4} \varepsilon_{\text{FB}} \frac{g_X}{q_*(x_f)}. \quad (4.78)$$

If the evolution of the Universe remains adiabatic after this transition then

$$n_{X0} = s_0 Y_\infty = 776.8 \left(\frac{q_{*0}}{3.91} \right) \frac{\epsilon_{\text{FB}} g_X}{q_*(x_f)} \Theta_{2.7}^3 \text{ cm}^{-3}. \quad (4.79)$$

If this particle is still relativistic today, then it behaves as the massless neutrino discussed in Section 4.2.1.6. If its mass is greater than the temperature of the current cosmic background, i.e. $m > 1.7 \times 10^{-4}$ eV, then the associated energy density is $\rho_{X0} = m_X s_0 Y_\infty$, so that using (4.140), its density is

$$\Omega_{X0} h^2 = m_X s_0 Y_\infty = 0.739 \left(\frac{q_{*0}}{3.91} \right) \left(\frac{m_X}{10 \text{ eV}} \right) \frac{\epsilon_{\text{FB}} g_X}{q_*(x_f)}. \quad (4.80)$$

The constraint $\Omega_X h^2 < 1$ implies that $m < 13.5 q_*(x_f) / \epsilon_{\text{FB}} g_X$ eV. The precise value of this constraint depends on the value of q_* at decoupling.

For instance, for slightly massive neutrinos, or for particles with a small mass that decouple at around 1 MeV, $q_*(x_f) = 10.75$, and we get

$$\Omega_{X0} h^2 = \frac{m_X}{91.5 \text{ eV}}, \quad (4.81)$$

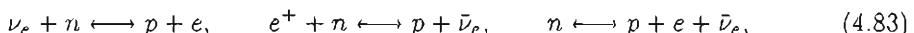
which gives a constraint on the particle mass. The relic density is inversely proportional to $q_*(x_f)$. A relic will thus be less abundant if it decouples early. For instance, if the particle decouples at around 300 GeV then $q_*(x_f) \sim 107$, so that

$$\Omega_{X0} h^2 = \frac{m_X}{910 \text{ eV}}. \quad (4.82)$$

So, a particle decoupling while $q_*(x_f) \gg 1$ will have a very small relic abundance and temperature compared to the photons of the cosmic background. These are usually called warm relics.

4.3 Primordial nucleosynthesis

For temperatures above or of the order of 100 MeV, the Universe is dominated by relativistic particles in equilibrium: electrons, positrons, neutrinos and photons, and the nuclear reactions can be activated. The contribution from neutrons and protons, which are then non-relativistic, is negligible in the mass budget. The weak interactions between neutrons, protons and leptons,



and the interactions between electrons and positrons,



maintain all these particles, together with non-relativistic baryons, in thermodynamical equilibrium.

4.3.1 Main stages of the mechanism

The synopsis of primordial nucleosynthesis (usually referred to as Big-Bang nucleosynthesis, BBN) can be decomposed into three main stages:

- (1) $T \gg 1 \text{ MeV}$, the components of the Universe are in thermodynamic equilibrium. Radiation dominates the Universe and $g_* = 10.75$. Thus, the neutron to proton ratio is approximatively given by its value at equilibrium

$$\left(\frac{n}{p}\right)_{\text{eq}} \sim \exp\left(-\frac{Q}{T}\right) \sim 1,$$

$Q = m_n - m_p = 1.293 \text{ MeV}$. The fraction of each atomic nucleus can then be obtained from purely thermodynamical considerations.

- (2) $T \sim 1 - 0.7 \text{ MeV}$, weak interactions can no longer maintain the equilibrium. Neutrinos decouple and the n/p ratio deviates from its equilibrium value. At the freeze-out temperature, $T_f \sim 0.8 \text{ MeV}$, it is of the order of

$$\left(\frac{n}{p}\right)_f \sim \exp\left(-\frac{Q}{T_f}\right) \sim \frac{1}{5}.$$

Free neutrons then decay into protons while atomic nuclei remain in thermodynamic equilibrium. In terms of time, this stage starts roughly round $t \sim 1 \text{ s}$ [see (4.21)].

- (3) $T \sim 0.7 - 0.05 \text{ MeV}$, the nuclear thermodynamic equilibrium can no longer be maintained. Electrons and positrons have annihilated each other and reheated the photon bath. g_* is then 3.36. The atomic nuclei are then formed by a series of two-body reactions. For that, deuterium needs to be synthesized ($p + n \rightarrow D + \gamma$), which is possible only when the radiation density is low enough so that the inverse reaction, the photodissociation of deuterium, is negligible. This determines the temperature T_{Nuc} obtained by $n_D/n_\gamma \sim 1$, i.e.

$$\eta^2 \exp\left(-\frac{B_D}{T_{\text{Nuc}}}\right) \sim 1.$$

Around T_{Nuc} , $(n/p)_{\text{Nuc}} = (n/p)_f \exp(-t_{\text{Nuc}}/\tau_n) \sim 1/7$. All neutrons are then bound in helium-4 nuclei, with abundance (in mass with respect to the total number of nucleons, $n_b = n + p$) of the order of

$$Y_p \sim \frac{2 \left(\frac{n}{p}\right)_{\text{Nuc}}}{1 + \left(\frac{n}{p}\right)_{\text{Nuc}}} \sim 0.25,$$

because the number density of helium-4 is $n/2$ so that it accounts for a mass $4 \times n/2 = 2n$ in nuclear units. The increase of the Coulomb barrier for nuclei of larger atomic mass together with the drop in the temperature and density (and hence in the probability of two-body reactions) associated with the absence of stable atomic

nuclei of mass $A = 5$ and $A = 8$ explain why primordial nucleosynthesis does not produce any nuclei beyond helium in any significant amount. BBN roughly stops when $T \sim 0.05$ MeV, which corresponds to $t \sim 530$ s [see (4.21)].

Key parameters

In this mechanism, departure from thermodynamic equilibrium is crucial, since all nucleons would otherwise be in the form of iron if the equilibrium was maintained during the entire nucleosynthesis. The abundances predicted by primordial nucleosynthesis depend mainly on a few parameters.

- g_* : the number of relativistic degrees of freedom. g_* determines the radiation density and thus the value of the freeze-out temperature T_f . This parameter depends on the number, N_ν , of neutrino families. The temperature T_f is roughly obtained from the Friedmann equation

$$G_F^2 T_f^5 \sim \sqrt{G_N g_*} T_f^2.$$

- τ_n : the neutron lifetime determines the ratio $(n/p)_{\text{Nuc}}$ at the beginning of nucleosynthesis from its freeze-out value. τ_n depends on various fundamental constants

$$\tau_n^{-1} = \frac{1.636}{2\pi^3} G_F^2 (1 + 3g_A^2) m_e^5 = (887 \pm 2 \text{ s})^{-1}.$$

- η : the number of baryons per photon $\eta \equiv n_b/n_\gamma \sim (5 - 6) \times 10^{-10}$ [see (4.142)]. This quantity is not directly measurable with a great accuracy as it is difficult to determine which fraction of dark matter is in baryonic form. BBN thus offers a way to measure this parameter.
- The fundamental constants: G_N and G_F are involved in the determination of the freeze-out temperature. The fine structure constant α determines Q and the reaction rates, since it enters the value of the Coulomb barriers.

This chapter will detail these main lines. To obtain precise results, one should integrate numerically the network of nuclear reactions in an expanding space-time. We refer to the original papers [1–3, 19, 20] and reviews [21–25] for more detailed explanations and to the numerical code [26] that is freely available [27].

4.3.2 Initial state

4.3.2.1 Thermodynamic nuclear equilibrium

At temperatures high compared to 1 MeV, all nuclei X_Z^A are kept in thermodynamic equilibrium by nuclear interactions. The density of these non-relativistic nuclei is then given by

$$n_A = g_A \left(\frac{m_A T}{2\pi} \right)^{3/2} e^{-(m_A - \mu_A)/T}. \quad (4.85)$$

As long as the reaction rate is greater than the expansion rate, the chemical equilibrium imposes the value of the chemical potential

$$\mu_A = Z\mu_p + (A - Z)\mu_n, \quad (4.86)$$

where μ_p and μ_n are the chemical potentials of protons and neutrons. Neglecting the mass difference $Q = m_n - m_p \simeq 1.293$ MeV in the pre-factors, the neutron and proton densities are

$$n_{(p/n)} = 2 \left(\frac{m_N T}{2\pi} \right)^{3/2} e^{-[m_{(p/n)} - \mu_{(p/n)}]/T}, \quad (4.87)$$

where $m_N = \frac{1}{2}(m_p + m_n)$.

Using (4.86), the exponential factor in the expression (4.85) can be expressed as

$$\exp \left(-\frac{m_A - \mu_A}{T} \right) = \left[\exp \left(\frac{\mu_p}{T} \right) \right]^Z \left[\exp \left(\frac{\mu_n}{T} \right) \right]^{A-Z} \exp \left(-\frac{m_A}{T} \right).$$

Expressing the first two factors as a function of the proton and neutron densities (4.87), we get

$$e^{-(m_A - \mu_A)/T} = 2^{-A} n_p^Z n_n^{A-Z} \left(\frac{2\pi}{m_N T} \right)^{3A/2} e^{B_A/T}, \quad (4.88)$$

where B_A is the nucleus binding energy

$$B_A = Zm_p + (A - Z)m_n - m_A. \quad (4.89)$$

The mass fraction of a nucleus with atomic number A with respect to the total nucleon mass is defined by

$$X_A \equiv \frac{An_A}{n_b}, \quad n_b = n_n + n_p + \sum An_A. \quad (4.90)$$

In terms of the parameter η defined by

$$\eta \equiv \frac{n_b}{n_\gamma}, \quad (4.91)$$

we finally get that the abundances of the atomic nuclei in thermodynamic equilibrium, using that $n_\gamma = 2\zeta(3)T^3/\pi^2$, are given by

$$X_A = F(A) \left(\frac{T}{m_N} \right)^{3(A-1)/2} \eta^{A-1} X_p^Z X_n^{A-Z} e^{B_A/T}, \quad (4.92)$$

$$F(A) \equiv g_A A^{5/2} \left[\zeta(3) \pi^{-1/2} 2^{(3A-5)/(2A-2)} \right]^{A-1}. \quad (4.93)$$

This illustrates the influence of the entropy on the abundance of the different nuclei. If $\eta \sim 1$, X_A^A is stable as soon as $T \sim B_A$ since at this temperature the formation of the nuclei, controlled by the factor $\exp(B_A/T)$, dominates over its destruction by photodissociation, controlled by the factor η^{A-1} . Now, if $\eta \ll 1$, the balance between formation and photodissociation is reached only when $\exp(B_A/T) \sim \eta^{A-1}$, that is at lower temperatures.

The temperature at which the abundance of the nucleus A becomes of order unity, which we denoted by T_A , can thus be estimated from (4.92). Neglecting the numerical factor $F(A)$ and with $X_p \sim X_n \sim 1$, we have

$$T_A \sim \frac{B_A(A-1)^{-1}}{-1.5 \ln(T_A/m_N) - \ln \eta}. \quad (4.94)$$

For the lightest elements, we obtain, with $\eta \sim 5 \times 10^{-10}$,

| | D | T | ^3He | ^4He |
|-------------|-------|------|---------------|---------------|
| B_A (MeV) | 2.22 | 6.92 | 7.72 | 28.3 |
| T_A (MeV) | 0.066 | 0.1 | 0.11 | 0.28 |

So, even at thermodynamic equilibrium, nucleosynthesis does not start before around 0.3 MeV. This is due to the very small value of η . Even if nuclei are energetically favoured, entropy acts in the opposite way and favours free neutrons and protons.

4.3.2.2 Neutron abundance

Since the equilibrium between protons and neutrons is maintained by weak interaction, (4.83) implies that

$$\mu_n + \mu_\nu = \mu_p + \mu_e. \quad (4.95)$$

We conclude from (4.87) that

$$\frac{n}{p} \equiv \frac{n_n}{n_p} = \exp \left(-\frac{Q}{T} + \frac{\mu_e - \mu_\nu}{T} \right). \quad (4.96)$$

As shown by (4.26), the electrical neutrality of the Universe implies that $\mu_e/T \sim 10^{-10} - 10^{-8}$. Neglecting μ_ν , we get that

$$\left(\frac{n}{p} \right)_{eq} = \exp \left(-\frac{Q}{T} \right). \quad (4.97)$$

For temperatures larger than 0.1 MeV, the study of the previous section teaches us that $X_A \ll 1$ so that $n_b = n_n + n_p$. Thus, the abundance of free neutrons in equilibrium is

$$X_{n,eq} = \left(1 + e^{Q/T} \right)^{-1}. \quad (4.98)$$

At large temperatures, we therefore have $X_n = X_p = \frac{1}{2}$ and $X_A \simeq 0$. For instance, for $T = 1$ MeV, deuterium has an abundance $X_2 \sim 6 \times 10^{-12}$.

4.3.3 Freeze-out of the weak interaction and neutron to proton ratio

Weak interactions maintain neutrons and protons in equilibrium. These interactions freeze-out at a temperature T_f determined by $H = \Gamma_w(T_f)$. For temperatures $T < Q$ and $T > m_e$ the reaction rate of the weak interactions is given by

$$\Gamma_w = \frac{7\pi}{60} (1 + 3g_A^2) G_F^2 T^5, \quad (4.99)$$

for the reaction $p + e \rightarrow \nu_e + n$. We obtain from (4.19) with $g_* = 10.75$ that

$$\frac{\Gamma}{H} \simeq \left(\frac{T}{0.8 \text{ MeV}} \right)^3, \quad (4.100)$$

so that the freeze-out temperature of the weak interactions is of the order of

$$T_f \sim 0.8 \text{ MeV}. \quad (4.101)$$

Thus, the freeze-out occurs around $t_f = 1.15$ s.

Applied to neutrons, the Boltzmann (4.56) can be re-expressed in terms of X_n as

$$\dot{X}_n = \lambda_{pn} (1 - X_n) - \lambda_{np} X_n. \quad (4.102)$$

The reaction rate $n \rightarrow p$, λ_{np} , is related to that of the reactions $p \rightarrow n$, λ_{pn} , by $\lambda_{np} = \lambda_{pn} \exp(Q/T)$ simply because the reactions (4.83) are reversible. In equilibrium, i.e. when $\dot{X}_n = 0$, the solution of this equation is given by (4.98).

We can first note that the evolution equation (4.102) has an integral solution

$$X_n(t) = \int_{t_i}^t I(t, t') \lambda_{pn}(t') dt' + I(t, t_i) X_n(t_i), \quad (4.103)$$

where the function I is defined by $I(t, t') = \exp \left\{ - \int_{t'}^t [\lambda_{pn}(u) + \lambda_{np}(u)] du \right\}$. Since at high temperatures the reaction rates are very large, $I(t, t_i)$ will be negligible if the time t_i is chosen to be far enough in the past, so that $X_n(t_i)$ plays no role and one can choose $t_i = 0$ (see, e.g., Ref. [20]). It follows that

$$X_n(t) = \int_0^t I(t, t') \lambda_{pn}(t') dt' \simeq X_{n,\text{eq}}(t) - \int_0^t I(t, t') \frac{d}{dt'} [X_{n,\text{eq}}(t')] dt'. \quad (4.104)$$

Since the reaction rates are large compared to their relative variation, it follows that

$$X_n(t) \simeq X_{n,\text{eq}}(t) - \frac{\dot{X}_{n,\text{eq}}}{\lambda_{pn}(t) + \lambda_{np}(t)} \simeq X_{n,\text{eq}} \left(1 + \frac{H}{\lambda_{pn} + \lambda_{np}} \frac{d \ln X_{n,\text{eq}}}{d \ln T} \right). \quad (4.105)$$

Thus, the neutron abundance follows its equilibrium value until T_f . When the weak interactions freeze-out, the neutron abundance is therefore

$$X_n(T_f) \simeq X_{n,\text{eq}}(T_f) = [1 + \exp(1.293/0.8)]^{-1} \sim 0.165 \sim \frac{1}{6}, \quad (4.106)$$

that is $(n/p)_f \sim 0.198 \sim 1/5$. The numerical coincidence $Q/T_f = \mathcal{O}(1)$ implies that X_n is neither equal to X_p nor very small. As a consequence, the quantity of helium synthesized will be of the same order of magnitude as the quantity of hydrogen. We stress here that this coincidence is unexpected since Q is determined by the strong and electromagnetic interactions, whereas T_f is fixed by the weak and gravitational interactions.

To evaluate the remaining abundances after the freeze-out, $X_n(t \gg t_f)$ one should use the explicit form of the reaction rates (see Ref. [22])

$$\begin{aligned}\lambda(n\nu_e \rightarrow pe) &= A \int_0^\infty dq_\nu q_\nu^2 q_e(E_\nu + Q)(1 - f_e)f_\nu, \\ \lambda(ne^+ \rightarrow p\bar{\nu}_e) &= A \int_0^\infty dq_e q_e^2 q_\nu(E_e + Q)(1 - f_\nu)f_e, \\ \lambda(n \rightarrow pe^- \bar{\nu}_e) &= A \int_0^{\sqrt{Q^2 - m_e^2}} dq_e q_e^2 q_\nu(E_e + Q)(1 - f_\nu)(1 - f_e).\end{aligned}\tag{4.107}$$

One can show (see Ref. [20] for details of this computation) that the leading part of these integrals comes from particles with energy large compared to the temperature during BBN. The Fermi-Dirac distributions can thus be approximated by Maxwell distributions, $f_e = [1 + \exp(E_e/T_e)]^{-1} \simeq \exp(-E_e/T_e)$. Since the Boltzmann factors are small in a diluted gas, quantum effects are negligible and the Pauli factors can be neglected, $1 - f \simeq 1$. The last approximation assumes $m_e = 0$ in the computation of the two first integrals, which are dominated by $E_{e,\nu} \gg m_e$, so that

$$\lambda(n\nu_e \rightarrow pe) = \lambda(ne^+ \rightarrow p\bar{\nu}_e) = AT^3(24T^2 + 12TQ + 2Q^2).$$

This approximation is precise at a level of 15% as long as $T > m_e$, when the reaction rates have become very weak. The last reaction rate gives a relation between the constant A and the neutron lifetime

$$\tau_n^{-1} = AQ^5 \left[\frac{\sqrt{1 - \tilde{m}^2}}{5} \left(\frac{1}{6} - \frac{3}{4}\tilde{m}^2 - \frac{2}{3}\tilde{m}^4 \right) + \frac{\tilde{m}^4}{4} \cosh^{-1} \left(\frac{1}{\tilde{m}} \right) \right] \simeq 0.0158AQ^5,$$

with $\tilde{m} \equiv m_e/Q$. Finally, the total reaction rate $\lambda_{np} \simeq 2\lambda(n\nu_e \rightarrow pe)$ takes the form

$$\lambda_{np} \simeq \frac{253}{\tau_n x^5} (12 + 6x + x^2),\tag{4.108}$$

with $x \equiv Q/T$, if we neglect the decay of neutrons during the freeze-out. Actually, this term only becomes important when $T < 0.13$ MeV, i.e. for $x > 10$. We recover that for $x \ll 1$, the reaction rate is given by $\lambda_{np} \sim 12 \times 253/\tau_n x^5 \sim G_F^2 T^5$, which justifies a posteriori our approximation in the determination of T_f .

The expression (4.108) allows us to compute the function $I(t, t')$ as

$$I(x, x') = \exp|K(x) - K(x')|,$$

with

$$\begin{aligned}K(x) &= b \int_x^\infty \frac{1 + e^{-u}}{u^4} (12 + 6u + u^2) du \\ &= b \left[\left(\frac{4}{x^3} + \frac{3}{x^2} + \frac{1}{x} \right) + \left(\frac{4}{x^3} + \frac{1}{x^2} \right) e^{-x} \right],\end{aligned}$$

where the parameter b is given by $b = 253\sqrt{45/4\pi^3 g_*} M_p / \tau_n Q^2 = 0.252$. From the expression (4.104), we can derive the evolution of the neutron abundance

$$X_n(x) = X_{n,\text{eq}}(x) + \int_0^x [X_{n,\text{eq}}(x')]^2 \exp[K(x) - K(x')] e^{x'} dx'. \quad (4.109)$$

Finally, we find that the neutron abundance freezes out at

$$X_n(\infty) = 0.150. \quad (4.110)$$

This value is reached at $T \sim 0.25$ MeV ($x \sim 5$), i.e. at $t \simeq 20$ s.

4.3.4 Abundances of the light elements

The previous computation does not take into account the change in g_* and the production of deuterium, two processes happening below 1 MeV.

Light elements will be formed by a series of nuclear reactions, summarized in Fig. (4.9). No significant abundance can be produced before the deuterium is formed since the densities are too low to allow for more than two-body reactions, such as $2n + 2p \rightarrow {}^4\text{He}$, to play any important role. Furthermore, the production of deuterium from two protons ($p + p \rightarrow D + e^+ + \nu_e$) involves the weak interaction, and is thus negligible.

Light isotopes cannot be synthesized before the temperature drops below 0.1 MeV. This is due to the fact that deuterium has a very weak binding energy and is easily photodissociated. Its synthesis only starts when the radiation density is low enough. Moreover, the reaction chain cannot continue before the deuterium is synthesized, this is often called the ‘deuterium bottleneck’.

4.3.4.1 Deuterium

The synthesis of deuterium is thus a key stage of primordial nucleosynthesis. It must be produced in sufficient quantity so that the heavy elements can then be synthesized.

As already mentioned, the synthesis of light isotopes only starts at around $T = T_{\text{nuc}} = 0.066$ MeV. At this temperature, the electrons and positrons are no longer relativistic and $g_* = 3.36$, so that (4.21) implies that $t_{\text{nuc}} = 303$ s. Since $n(T_{\text{nuc}}) = n(T_f) \exp(-t_{\text{nuc}}/\tau_n)$ and $p(T_{\text{nuc}}) = p(T_f) + n(T_f)[1 - \exp(-t_{\text{nuc}}/\tau_n)]$, which imply that $(n/p)(T_{\text{nuc}}) \sim 0.133$, we have

$$X_n(T_{\text{nuc}}) \simeq 0.117. \quad (4.111)$$

The reaction rate per neutron of the deuterium-forming reaction ($n + p \rightarrow D + \gamma$) is

$$\lambda_D = 4.55 \times 10^{-20} n_p \text{ cm}^3 \cdot \text{s}^{-1}. \quad (4.112)$$

It can be shown that this reaction freezes out well after the end of nucleosynthesis so that the deuterium abundance ($g_D = 3$) is given by its equilibrium value (4.85)

$$\frac{X_D}{X_n X_p} = \frac{24\zeta(3)}{\sqrt{\pi}} \eta \left(\frac{T}{m_N} \right)^{3/2} \exp \left(\frac{B_D}{T} \right). \quad (4.113)$$

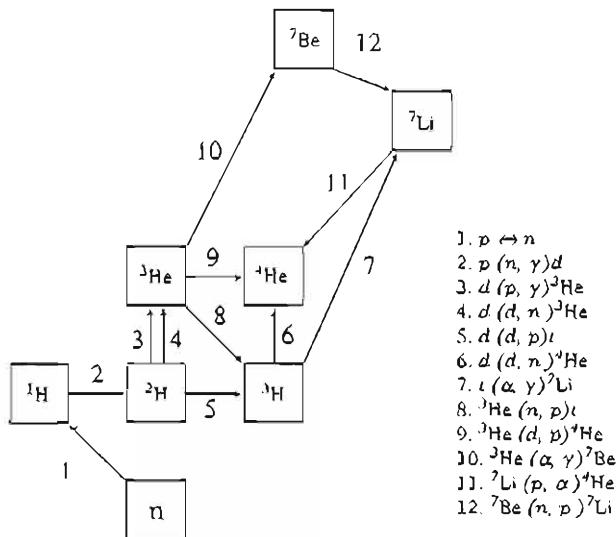


Fig. 4.9 Simplified network of nuclear reactions involved in the synthesis of light nuclei during BBN.

4.3.4.2 Helium-4

Since the binding energy of helium-4 is larger than that of deuterium, helium is favoured by the Boltzmann factor $\exp(B/T)$ in comparison to deuterium. As can be seen from Fig. 4.10, the growth in the deuterium abundance is followed by that of helium-4. Heavier nuclei are not synthesized in significant quantities mainly for two reasons: (1) the absence of stable nuclei with $A = 5$ or $A = 8$, so that the production chains for the reactions $n+{}^4He$, $p+{}^4He$, ${}^4He+{}^4He$ are blocked (the triple- α process, that produces carbon-12 in stars, cannot be efficient in the primordial Universe); (2) the important Coulomb barrier for reactions as ${}^3H+{}^4He \rightarrow {}^7Li + \gamma$ and ${}^3He+{}^4He \rightarrow {}^7Be + \gamma$.

At T_{nuc} , all free neutrons are bound in helium-4 nuclei. Since helium-4 has two neutrons, its primordial abundance is therefore

$$Y_p = \frac{2X_n(T_{nuc})}{1 + X_n(T_{nuc})} \simeq 0.21. \quad (4.114)$$

A more precise estimate [20] of T_{nuc} gives $T_{nuc} = 0.086$ MeV, which implies $t_{nuc} \simeq 178.5$ s and hence $Y_p \sim 0.238$.

This analytical estimate gives the correct orders of magnitude and predictions with a precision of the order of 1% as compared with what can be obtained from numerical integrations.

Let us recall that our estimate assumed $\mu_\nu = 0$ to determine the ratio (4.97). If this is not the case, the constant parameter $\xi \equiv \mu_\nu/T$ can be introduced. When this parameter is small, its main effect is to modify the reaction $n + \nu_e \rightarrow p + e$, so

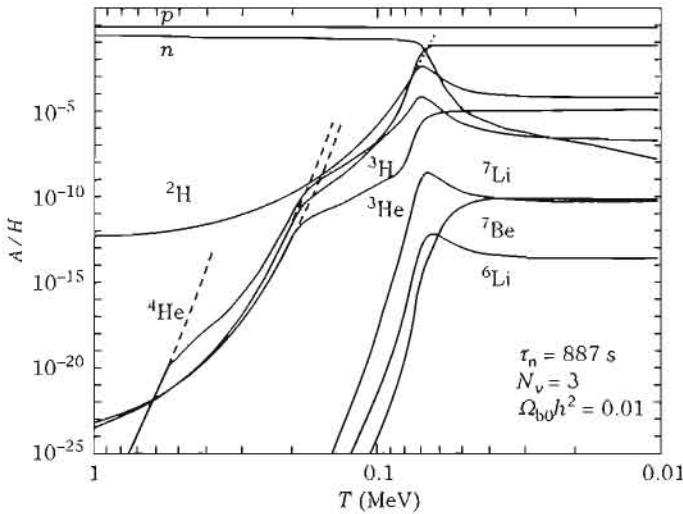


Fig. 4.10 Evolution of the abundances of light elements during primordial nucleosynthesis. Dashed lines represent abundances in thermodynamic equilibrium.

that $\lambda_{np} = \lambda_{pn} \exp(Q/T + \xi)$. This changes the freeze-out neutron abundance (4.110) to $X_n(\infty) = X_n(\infty, \xi = 0) \exp(-\xi)$.

The results from numerical simulations can be summarized, in the interval $3 \times 10^{-10} < \eta < 10^{-9}$, by

$$Y_p = 0.245 + 0.014(N_\nu - 3) + 0.0002(\tau_n - 887 \text{ s}) + 0.009 \ln \left(\frac{\eta}{5 \times 10^{-10}} \right) - 0.25\xi, \quad (4.115)$$

which gives the sensitivity to the most critical parameters (see Ref. [28] for a more detailed study).

4.3.4.3 Other elements

At the end of nucleosynthesis, at around $t \sim 1000$ s, traces of deuterium ($X_D \sim 10^{-4}$), helium-3 ($X_{^3\text{He}} \sim 10^{-4}$) and lithium ($X_{^7\text{Li}} \sim 10^{-9}$) are still present. The abundance of helium-3 includes some of the tritium that disappears by β decay, and similarly, the one of lithium-7 includes beryllium-7. The production of elements beyond $A = 7$, such as boron, is negligible.

Unlike helium-4, the abundances of these nuclei are very sensitive to the value of η (see Fig. 4.11). The abundances of deuterium and helium-3 decrease with η . The lithium-7 can be produced by two channels: (1) ${}^4\text{He} + {}^3\text{H} \rightarrow {}^7\text{Li} + \gamma$ that competes with ${}^7\text{Li} + p \rightarrow {}^4\text{He} + {}^4\text{He}$ for $\eta < 3 \times 10^{-10}$ and (2) ${}^4\text{He} + {}^3\text{He} \rightarrow {}^7\text{Be} + \gamma$, followed by a β decay for $\eta > 3 \times 10^{-10}$. The transition between both channels explains the shape of the lithium-7 abundance curve (Fig. 4.11).

We can show [23] that in the band $10^{-10} < \eta < 10^{-9}$, the primordial abundances of the light nuclei are given by

$$\left(\frac{D}{H}\right)_p = 3.6 \times 10^{-5 \pm 0.06} \eta_{5.5}^{-1.6}, \quad (4.116)$$

$$\left(\frac{^3\text{He}}{H}\right)_p = 1.2 \times 10^{-5 \pm 0.06} \eta_{5.5}^{-0.63}, \quad (4.117)$$

$$\left(\frac{^7\text{Li}}{H}\right)_p = 1.2 \times 10^{-11 \pm 0.2} (\eta_{5.5}^{-2.38} + 21.7 \eta_{5.5}^{2.38}), \quad (4.118)$$

with $\eta_{5.5} \equiv \eta/5.5 \times 10^{-10}$.

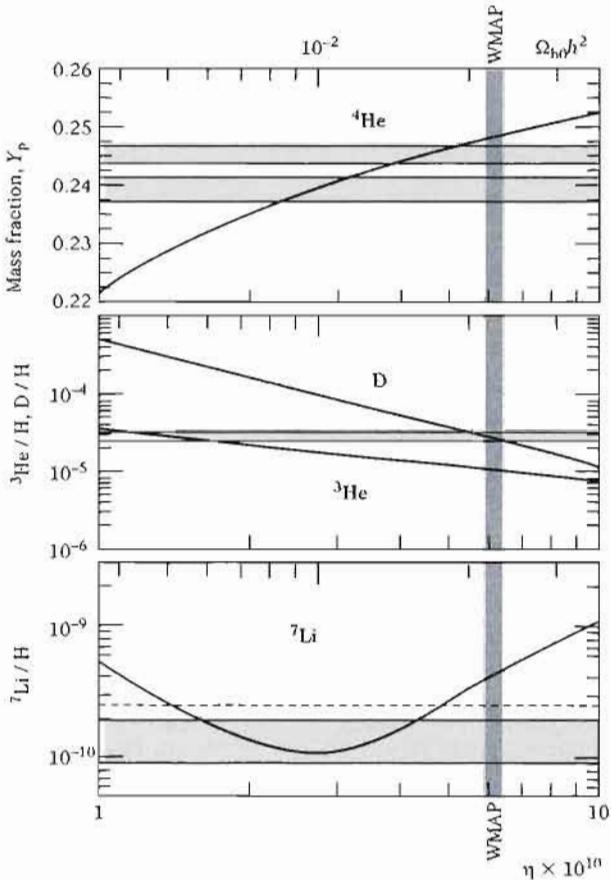


Fig. 4.11 Helium-4, deuterium, helium-3 and lithium-7 abundances as a function of the baryon to photon ratio, η , or equivalently of the baryon density, $\Omega_{b0}h^2$. The hatched areas correspond to the spectroscopic observations discussed in the text: Refs. [29–31] for helium-4, Ref. [32] for deuterium, and Ref. [33] for lithium-7. The vertical band represents the constraint on η obtained from the WMAP satellite [34]. From Ref. [35].

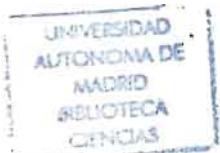
4.3.5 Observational status

The comparison between the observed and predicted abundances is difficult, mainly because the primordial abundances can be modified by various nuclear processes during the evolution of the Universe. Moreover, these modifications are different for each nucleus. For instance, the abundance of helium-4, which is very stable, increases due to the stellar production, whereas the deuterium is much less stable and can only be destroyed but never be synthesized in stars. Helium-3 and lithium-7 have more complex evolutions as they can be both synthesized and destroyed. Astronomers therefore try to measure the abundances in environments as primordial as possible.

4.3.5.1 Observations

We briefly review the status of the light-nuclei abundance observations. For a more detailed review on these topics, see Ref. [24].

- *Helium-4*: is detected through its emission lines in various astrophysical environments (planetary atmospheres, young stars, planetary nebulae), which could be galactic, extragalactic or intergalactic. In order to eliminate the production of stellar helium, its abundance is correlated to that of other elements such as nitrogen and oxygen. The best system to carry out these studies seems to be the HII regions of compact blue galaxies. Recent measurements from various groups give $Y_p = 0.2443 \pm 0.0015$ [29], $Y_p = 0.2391 \pm 0.0020$ [30] and $Y_p = 0.249 \pm 0.004$ [36]. The recent analysis of 82 HII regions performed in 76 compact galaxies has given [31] $Y_p = 0.2421 \pm 0.0021$.
- *Deuterium*: is easily destroyed in stars as soon as the temperature increases above 6×10^5 K. Any measurement of D/H is therefore a lower bound, giving an upper bound on η . Its abundance could have been reduced by a factor estimated from 2 to 10. Its spectral lines have actually never been detected in any star, which implies that $D/H < 10^{-6}$ in stellar atmospheres. They can be detected in giant planets, in local interstellar environments and in the absorption spectrum of low-metallicity clouds on the line of sight of quasars. The finest measurements in absorption systems lead [32] to $D/H = 2.78_{-0.38}^{+0.44} \times 10^{-5}$. Conservatively, local measurements within the interstellar regions give the lower bound [37] $D/H = (1.5 \pm 0.1) \times 10^{-5}$, while the absence of any detection in systems with high redshift fixes the upper bound [32] $D/H < 6.7 \times 10^{-5}$.
- *Lithium-7*: the system that seems to enable the best measurements of lithium-7 is the stellar halo of our Galaxy. This halo is mainly composed of stars from populations I and II and thus has a very low metallicity. Extrapolating to zero metallicity, we get a primordial abundance $Li/H = 1.23_{-0.32}^{+0.68} \times 10^{-10}$ [33].
- *Helium-3*: has only been observed in HII regions of the galactic disk (thus in a relatively high metallicity environment). In these regions, its abundance is ${}^3He/H = (1 - 2) \times 10^{-5}$ [38] although it is difficult to track back to its primordial value since this isotope has a complex stellar history and can be both produced and destroyed during the galactic evolution. Helium-3 is thus unlikely to provide a precise cosmological test today. At the moment, it can be mainly



used to constrain the astrophysical mechanisms in which it is produced or destroyed. However, by comparing the primordial abundance of ^3He with the disk abundance, we can conclude that it has not evolved much in 14 billion years.

4.3.5.2 Constraints on cosmological models

Primordial nucleosynthesis allows us first to measure the ratio η . If only spectroscopic data are considered and assuming $N_\nu = 3$, then

$$\eta = (5.15 \pm 1.75) \times 10^{-10}, \quad (4.119)$$

which corresponds to $\Omega_{b0}h^2 = 0.018 \pm 0.006$.

Current data allow us to impose constraints on parameters such as N_ν , the number of neutrino families with mass lower than 1 MeV, and ξ , but also to test the value of some fundamental constants such as the gravitational constant, or the fine structure constant (see Ref. [39]). Primordial nucleosynthesis is thus a test not only of the Big-Bang model, but also of nuclear physics and of general astrophysics. The agreement between theory and observations makes it one of the pillars of the Big-Bang model.

4.3.5.3 A recent crisis?

For a long time, primordial nucleosynthesis and the spectroscopic determination of light-element abundances were the only reliable ways to quantify the baryon content of the Universe. Recently, $\Omega_{b0}h^2$ has been determined with an extreme precision [34] thanks to the analysis of the cosmic microwave background anisotropies (see Chapter 6).

$$\Omega_{b0}h^2 = 0.0224 \pm 0.0009, \quad (4.120)$$

which corresponds to $\eta = (6.14 \pm 0.25) \times 10^{-10}$. Using this value for the baryon density, we can then deduce the expected abundances [35]:

| y_b | D/H | $^7\text{Li}/\text{H}$ | $^3\text{He}/\text{H}$ | $^6\text{Li}/\text{H}$ |
|---------------------|---------------------------------------|--|----------------------------------|----------------------------------|
| 0.2479 ± 0.0004 | $2.80^{+0.19}_{-0.17} \times 10^{-5}$ | $4.15^{+0.49}_{-0.45} \times 10^{-10}$ | $(1.04 \pm 0.04) \times 10^{-5}$ | $(1.2 \pm 0.08) \times 10^{-14}$ |

This has opened up a reanalysis of the nuclear reaction rates involved during nucleosynthesis (Fig. 4.9). The situation is summarized in Figs. 4.11.

4.4 The cosmic microwave background radiation

As long as the temperature of the Universe remains large compared to the hydrogen ionization energy, matter is ionized and photons are then strongly coupled to electrons through Compton scattering. At lower temperatures, the formation of neutral atoms is thermodynamically favoured for matter. Compton scattering is then no longer efficient and radiation decouples from matter to give rise to a fossil radiation: the cosmic microwave background (CMB).

The temperature of this cosmic background was predicted by Alpher and Herman [3] in 1948, following the arguments developed in the same year by Gamow [2]. This prediction was confirmed in 1964 by Penzias and Wilson [40] who discovered

an unexplained noise signal, identical in every direction. This signal was immediately interpreted by Dicke *et al.* [41] as being the cosmic microwave background radiation at a temperature of 3.5 ± 1 K.

In this section, we first describe the recombination and decoupling mechanisms before describing the properties of the isotropic component of the cosmic background as revealed by the COBE satellite in 1987; in particular, we describe the constraints on the possible distortions of the Planck spectrum. For more details on this subject, we recommend the reviews [42–44].

4.4.1 Recombination

4.4.1.1 Prediction of the existence of the cosmic microwave background

To start with, let us briefly review the analysis of Alpher and Herman who predicted the temperature of the cosmic background in a surprising way, based uniquely on an estimate of the helium abundance and on the baryon density.

To explain a helium-4 abundance of approximatively 25%, primordial nucleosynthesis must have happened at around 10^9 K. At this time, the baryon density is of the order of $n_b \sim 10^{18} \text{ cm}^{-3}$. Today, it is of order $n_{b0} \sim 10^{-7} \text{ cm}^{-3}$, from which we deduce that, at the time of nucleosynthesis, the redshift is

$$1 + z_{\text{BBN}} = \left(\frac{n_b}{n_{b0}} \right)^{1/3} \sim 2 \times 10^8.$$

Since T scales as $(1+z)$, the current temperature of the photon bath is thus around

$$T_{\text{CMB}} = \frac{T_{\text{BBN}}}{1 + z_{\text{BBN}}} \sim 5 \text{ K}.$$

4.4.1.2 Recombination and decoupling

As long as the photoionization reaction



is able to maintain the equilibrium, the relative abundances of the electrons, protons and hydrogen will be fixed by (4.60). Since the Universe is electrically neutral, one has $n_e = n_p$. For simplicity, we introduce the ionization fraction

$$X_e = \frac{n_e}{n_p + n_H}, \quad (4.122)$$

where the denominator represents the total number of hydrogen nuclei, $n_b = n_p + n_H$ [$n_e = n_p = X_e n_b$, $n_H = (1 - X_e) n_b$]. In this particular case, (4.60) is known as the Saha equation. Once the equilibrium quantities are computed, it implies that

$$\frac{X_e^2}{1 - X_e} = \left(\frac{m_e T}{2\pi} \right)^{3/2} \frac{e^{-E_I/T}}{n_b}, \quad (4.123)$$

where $E_I = m_e + m_p - m_H = 13.6$ eV is the hydrogen ionization energy. The photon temperature is $T = 2.725(1+z)$ K = $2.348 \times 10^{-4}(1+z)$ eV and the baryon density $n_b = \eta n_{\gamma 0}(1+z)^3 \text{ cm}^{-3}$ so that the previous equation becomes

$$\log \left(\frac{X_e^2}{1 - X_e} \right) = 20.98 - \log \left[\Omega_{b0} h^2 (1+z)^{3/2} \right] - \frac{25163}{1+z}. \quad (4.124)$$

Notice first that when $T \sim E_1$, the right-hand side of (4.123) is of order 10^{15} , so that $X_e(T \sim E_1) \sim 1$. Recombination only happens for $T \ll E_1$. To see that, we deduce from (4.124) that if recombination is defined by $X_e = 0.5, 0.1, 0.01$ then $z_{\text{rec}} = 1376, 1258, 1138$, respectively. Thus, X_e varies abruptly between $z = 1400$ and $z = 1200$ and the recombination can be estimated to occur at a temperature between 3100 and 3800 K. Note that (4.123) implies that

$$\ln \frac{X_e^2}{1 - X_e} = \frac{3}{2} \ln \left(\frac{m_e T}{2\pi} \right) - \frac{E_1}{T} - \ln \left[\eta \frac{2}{\pi^2} \zeta(3) T^3 \right],$$

from which we deduce that

$$\frac{E_1}{T} = \frac{3}{2} \ln \left(\frac{m_e}{2\pi T} \right) - \ln \eta - \ln \left[\frac{2}{\pi^2} \zeta(3) \frac{X_e^2}{1 - X_e} \right].$$

This gives a rough estimate of the recombination temperature since the last term can be neglected. It gives $T \sim 3500$ K.

The electron density varies rapidly at the time of recombination, thus the reaction rate $\Gamma_T = n_e \sigma_T$, with σ_T the Thompson scattering cross-section, drops off rapidly so that the reaction (4.121) freezes out and the photons decouple soon after. An estimate of the decoupling time, t_{dec} , can be obtained by the requirement $\Gamma_T(t_{\text{dec}}) = H(t_{\text{dec}})$. The reaction rate Γ_T takes the form

$$\Gamma_T = n_e \sigma_T = 1.495 \times 10^{-31} X_e \left(\frac{\Omega_{b0} h^2}{0.02} \right) (1+z)^3 \text{ cm}^{-1}. \quad (4.125)$$

Only matter and radiation contribute significantly to the Hubble expansion rate so that

$$H^2 = \Omega_{m0} H_0^2 (1+z)^3 \left(1 + \frac{1+z}{1+z_{\text{eq}}} \right). \quad (4.126)$$

Thus, we conclude that the decoupling redshift is solution of

$$(1+z_{\text{dec}})^{3/2} = \frac{280.01}{X_e(\infty)} \left(\frac{\Omega_{b0} h^2}{0.02} \right)^{-1} \left(\frac{\Omega_{m0} h^2}{0.15} \right)^{1/2} \sqrt{1 + \frac{1+z_{\text{dec}}}{1+z_{\text{eq}}}}. \quad (4.127)$$

$X_e(\infty)$ is the residual electron fraction, once recombination has ended. We can estimate (see the next section) that $X_e(\infty) \sim 7 \times 10^{-3}$.

4.4.1.3 Recombination dynamics

To describe the recombination, and to determine $X_e(\infty)$, one should solve the Boltzmann equation. A complete treatment, taking the hydrogen and helium contributions into account, is described in Refs. [45, 46]. Actually, the helium recombination ends long before that of hydrogen begins. For simplicity, we will neglect the helium contribution.

Consider the Boltzmann equation (4.58) for the electron density. Using $n_e = n_b X_e$ and the fact that $n_b a^3$ is constant, the term $\dot{n}_e + 3Hn_e$ can be written as $n_b \dot{X}_e$. Then, using that $n_H = n_b(1 - X_e)$ and $\bar{n}_e \bar{n}_p / \bar{n}_H = (m_e T / 2\pi)^{3/2} \exp(-E_I/T)$, it reduces to

$$\dot{X}_e = C_r \left[\beta(1 - X_e) - \alpha^{(2)} n_b X_e^2 \right], \quad (4.128)$$

with

$$\beta = \left(\frac{m_e T}{2\pi} \right)^{3/2} \alpha^{(2)} \exp \left(-\frac{E_I}{T} \right), \quad \alpha^{(2)} = \langle \sigma_T v \rangle. \quad (4.129)$$

These parameters are given in terms of the temperature by

$$\alpha^{(2)} = \frac{64\sqrt{\pi}}{\sqrt{27}} \frac{e^4}{m_e^2} \left(\frac{T}{E_I} \right)^{-1/2} \phi_2(T), \quad \phi_2(T) \simeq 0.448 \ln \left(\frac{E_I}{T} \right). \quad (4.130)$$

The expression for ϕ_2 is a good approximation, better than 1% for $T < 6000$ K. The coefficient C_r should in principle be equal to unity.

This description is actually too naive [45]. When an electron is captured by a proton, a hydrogen atom is formed in an excited state. This atom returns to a lower excited state by emitting a series of resonant Lyman- α photons. To reach the ground state, one of the photons must have at least the transition energy $2s \rightarrow 1s$. These photons have an important cross-section and thus a high probability to be absorbed by surrounding hydrogen atoms, exciting them to a state from which they can easily be ionized. Thus, this process does not lead to any net change in the hydrogen density.

It has been shown by Peebles [45] that only transitions from the level $n = 2$ can lead to significant changes in the hydrogen abundance. The fine structure of this level therefore plays a central role during recombination. Two leading effects should be taken into account.

- The transition $2s \rightarrow 1s$ is forbidden to first order. From conservation of energy and angular momentum, two electrons should be emitted during this transition. This process is slow and its reaction rate is $\Lambda_{2s} = 8.227 s^{-1}$. This is the leading mechanism and the reaction rate of recombination is very different from the one predicted by the Saha equation.
- The transition $2p \rightarrow 1s$ produces some Lyman- α photons. The frequencies of these photons are redshifted by the cosmic expansion. Thus, their energy gets lower and they are no longer resonant, this lowers the probability to reionize a hydrogen atom.

These two effects are equivalent to fixing the value of the reduction coefficient, C_r , to

$$C_r = \frac{\Lambda_\alpha + \Lambda_{2s}}{\Lambda_\alpha + \Lambda_{2s} + \beta^{(2)}} \quad (4.131)$$

with

$$\beta^{(2)} = \beta \exp \left(\frac{\nu_\alpha}{T} \right), \quad \Lambda_\alpha = \frac{8H}{\lambda_\alpha^3 n_{1s}}, \quad \lambda_\alpha = \frac{4}{3E_I} = 1.216 \times 10^{-5} \text{ cm}^{-1}.$$

For $T < 10^5$ K the hydrogen abundance in the state $1s$ can be replaced by $n_{1s} = (1 - X_e)n_b$. The $\beta^{(2)}$ coefficient describes the ionization, whereas $\Lambda_\alpha + \Lambda_{2s}$ is the total

reaction rate of both dominant processes. A more complete study, with details of all the processes is presented in Ref. [46]. Table 4.6 compares the approximation from the Saha equation with the integration of (4.128) for a flat Universe with $\Omega_{b0}h^2 = 1$.

Table 4.6 Comparison between the approximation from the Saha equation and the integration of (4.128) for a flat Universe with $\Omega_{b0}h^2 = 1$.

| z | 1667 | 1480 | 1296 | 1111 | 926 | 741 |
|-----------|-----------------------|----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Saha | 9.43×10^{-1} | 3.8×10^{-1} | 2.6×10^{-2} | 8.8×10^{-4} | 5.28×10^{-6} | 2.4×10^{-9} |
| (4.128) | 9.2×10^{-1} | 4.0×10^{-1} | 7.2×10^{-2} | 9.8×10^{-3} | 9.2×10^{-4} | 1.23×10^{-4} |
| Ref. [46] | 9.14×10^{-1} | 4.0×10^{-1} | 7.15×10^{-2} | 9.68×10^{-3} | 9.01×10^{-4} | 1.2×10^{-4} |

A detailed analysis of (4.128) shows [46] that for $1500 > z > 800$, it reduces to

$$\frac{dX_e}{dz} = -61.72 \frac{\Omega_{b0}h^2}{(\Omega_{m0}h^2)^{1/2}} f(z) \left[1 + 2.26 \times 10^4 z e^{-14486/(z\Theta_{2.7})} \right]^{-1} X_e^2, \quad (4.132)$$

with $f(z) = (1 + 1.45z/z_{eq})^{-1/2}$, if we only keep the dominant contribution. A good approximation of the solution of this equation is

$$X_e(z) = 2.74 \times 10^{-3} e^{14.486(1-\Theta_{2.7}^{-1})} \frac{(\Omega_{m0}h^2)^{1/2}}{\Omega_{b0}h^2} \left(\frac{z}{1000} \right)^{12.75}, \quad (4.133)$$

as long Ω_{b0} and Ω_{m0} are both between 0.05 and 1.

4.4.1.4 Last scattering surface

The optical depth can be computed using the ionization fraction as a function of the redshift around decoupling,

$$\tau = \int n_e X_e \sigma_T d\chi \approx 0.42 e^{14.486(1-\Theta_{2.7}^{-1})} \left(\frac{z}{1000} \right)^{14.25}, \quad (4.134)$$

where the Ω_{m0} dependence disappears in an obvious way. More interestingly, τ is also independent of Ω_{b0} . The optical width varies rapidly around $z \sim 1000$ so that the visibility function $g(z) = \exp(-\tau) d\tau/dz$, which determines the probability for a photon to be scattered between z and $z + dz$, is a highly peaked function (see Fig. 4.12). Its maximum defines the decoupling time, $z_{dec} \approx 1057.3$. This redshift defines the time at which the CMB photons last scatter; the Universe then rapidly becomes transparent and these photons can propagate freely in all directions. The Universe is then embedded in this homogeneous and isotropic radiation. The instant when the photons last interact is a space-like hypersurface, called the *last scattering surface*. Some of these relic photons can be observed. They come from the intersection of the last scattering surface with our past light cone. It is thus a 2-dimensional sphere, centred around us, with comoving radius $\chi(z_{dec})$. This sphere is not infinitely thin. If we define its width as the zone where the visibility function is halved, we get $\Delta z_{dec} = 185.7$. As can be seen in (4.133) and (4.134), the value of z_{dec} is not

very sensitive to $\Omega_{b0}h^2$ and $\Omega_{m0}h^2$ [it is only independent of these parameters in the approximation of (4.132)]. From the analysis [34] of the WMAP satellite, we can conclude that

$$z_{\text{dec}} = 1089 \pm 1, \quad \Delta z_{\text{dec}} = 195 \pm 2, \quad (4.135)$$

which is in good agreement with our analytical approximation.

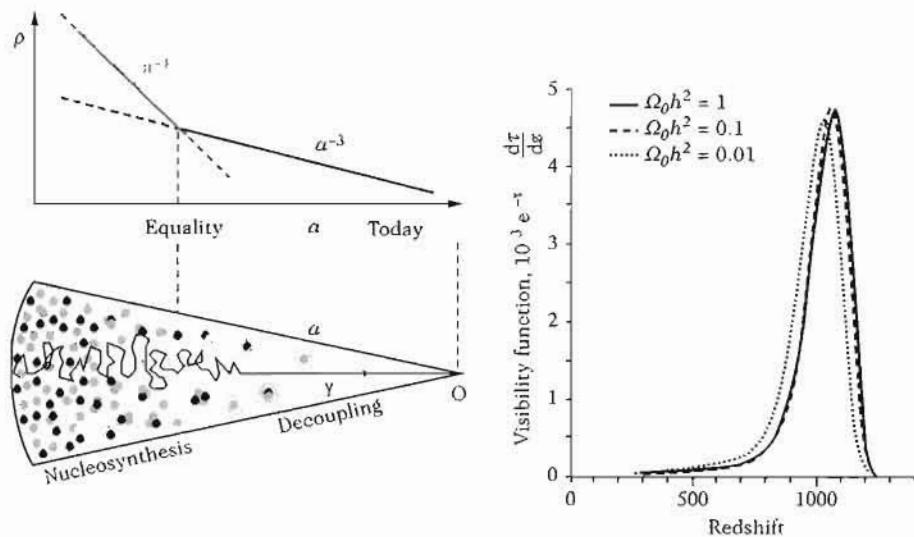


Fig. 4.12 (left): During the expansion of the Universe, the density, ρ , decreases with time. Slightly after recombination, the free-electron density rapidly decreases and the Universe becomes transparent. (right): The visibility function is highly peaked, which allows us to define the last scattering surface.

4.4.2 Properties of the cosmic microwave background

In Big-Bang models, the energy injected into radiation between the electron–positron annihilation at a redshift of order $z \sim 10^9$ and $z = 10 - 20$ is very small. The fossil radiation emitted around $z = 10^3$ must hence have a spectrum very close to a Planck distribution. We had to wait until the COBE satellite [42] to check this prediction with accuracy. In this section, we detail the observed properties of this radiation.

4.4.2.1 Observed temperature and spectrum

The temperature of the cosmic microwave background, defined as the average of the temperature of the whole sky

$$T_0 \equiv T_{\gamma 0} = \frac{1}{4\pi} \int T(\theta, \varphi) \sin \theta d\theta d\varphi,$$

has been measured with precision by the FIRAS experiment on board the COBE satellite [47]

$$T_0 = 2.725 \pm 0.001 \text{ K}$$

at 2σ . We define the dimensionless parameter

$$\Theta_{2.7} \equiv \frac{T_0}{2.725} \text{ K.} \quad (4.137)$$

This temperature corresponds to an energy of $E_{\gamma 0} = 2.345 \times 10^{-4} \text{ eV}$.

The observed spectrum is compatible with a blackbody spectrum

$$I(\nu) \propto \frac{\nu^3}{e^{\nu/T} - 1}. \quad (4.138)$$

Figs. 4.13 and 4.14 compare the measurements of the spectrum of the cosmic microwave background to that of a blackbody at 2.725 K.

The fact that this spectrum is so close to a blackbody proves that the fossil radiation could have been thermalized, mainly thanks to interactions with electrons. However, for redshifts lower than $z \sim 10^6$, the fossil radiation do not have time to be thermalized. Any energy injection at lower redshift would induce distortions in the Planck spectrum and can thus be constrained from these observations (see Fig. 4.13).

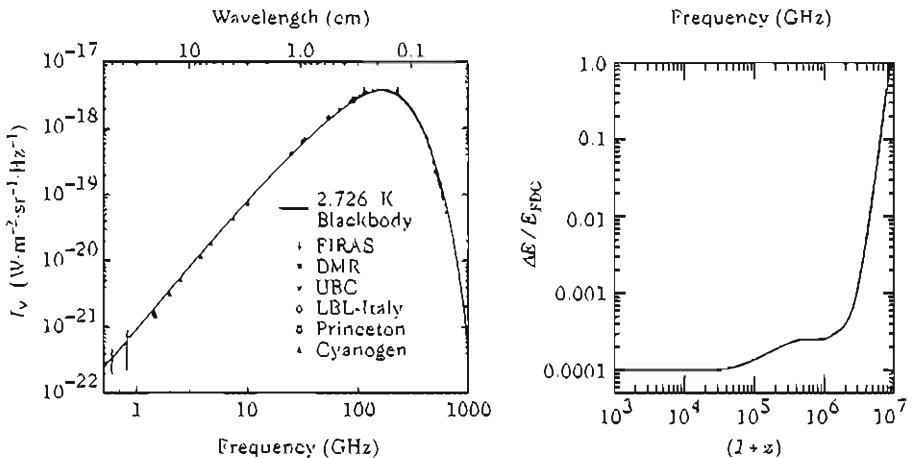


Fig. 4.13 (left): Comparison between the measurements of the cosmic microwave background spectrum to a blackbody at 2.725 K. (right): Upper bound on the injected energy compatible with the constraints from FIRAS on the distortions of the cosmic background spectrum. From Ref. [48].

CMB photons have a Planck distribution at a temperature T_0 . From the expressions (4.11) and (4.12), we can conclude ($g_\gamma = 2$) that

$$\rho_{\gamma 0} = 410.44 \Theta_{2.7}^3 \text{ cm}^{-3} \quad \rho_{\gamma 0} = 4.6408 \times 10^{-34} \Theta_{2.7}^4 \text{ g} \cdot \text{cm}^{-3}. \quad (4.139)$$

This implies that $\rho_{\gamma 0} = 0.26032 \text{ eV} \cdot \text{cm}^{-3} = 4.96 \times 10^{-13} \text{ eV}^4$. We then deduce from (4.29) that the current entropy density is

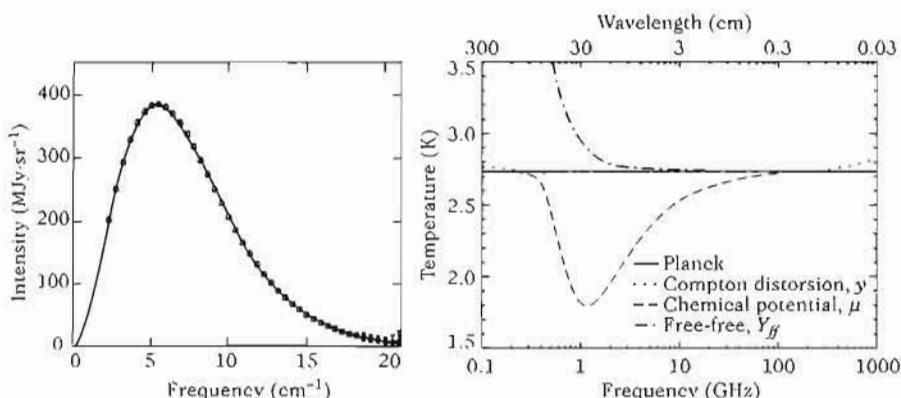


Fig. 4.14 (left): Spectrum of the fossil radiation measured by the COBE satellite, the error bars are multiplied by 200. (right): Different kinds of distortions of the blackbody spectrum. None of these distortions has been observed. From Ref. [48].

$$s_0 = 2889.2 \left(\frac{q_*}{3.91} \right) \Theta_{2.7}^3 \text{ cm}^{-3} = 7.039 \left(\frac{q_*}{3.91} \right) n_{\gamma 0}, \quad (4.140)$$

and that the photon density parameter is

$$\Omega_{\gamma 0} h^2 = 2.4697 \times 10^{-5} \Theta_{2.7}^4. \quad (4.141)$$

Using $n_{b0} = \Omega_{b0} \rho_{crit} / m_N \sim 0.11 (\Omega_{b0} / 0.02) \text{ m}^{-3}$ we deduce that

$$\eta = 5.514 \times 10^{-10} \left(\frac{\Omega_{b0} h^2}{0.02} \right) \Theta_{2.7}^{-3}. \quad (4.142)$$

4.4.2.2 Distortions of the Planck spectrum

Three kinds of distortions can affect the spectrum of the cosmic background.

- At redshifts lower than z_{dec} , the cosmic plasma can be heated by the injection of energy with temperatures higher than the one of the CMB photons. The Compton scattering of hot electrons then tends to shift the spectrum towards higher energies, while maintaining the number of photons constant, so that the low frequencies (Rayleigh-Jeans part) get depleted in favour of the higher frequencies (Wien part).

This distortion is called *Compton distortion* and is characterized by a unique parameter

$$y = \int \sigma_T n_e \frac{(T_e - T_\gamma)}{m_e c^2} dt, \quad (4.143)$$

which represents the integrated Compton optical depth.

- For redshifts between $z \sim 2 \times 10^6$ and $z \sim 10^5$, the cosmic microwave background can be thermalized through Compton scattering, but the thermalization of the energy between the plasma and the photons happens too late for the spectrum to

have time to relax into a Planck spectrum. The spectrum is then deformed to a Bose-Einstein spectrum with non-vanishing chemical potential

$$J(\nu) \propto \frac{\nu^3}{e^{(\nu - \mu)/T} - 1}. \quad (4.144)$$

This difference can be characterized by the dimensionless parameter

$$\tilde{\mu} = \frac{\mu}{T_\gamma}. \quad (4.145)$$

3. Finally, there can be some so-called free-free distortions, mainly coming from the emission of photons by electrons that scatter on charged particles. The effect on the cosmic background is then proportional to the parameter Y_F defined as

$$Y_F = \int \kappa \left(1 - \frac{T_\gamma}{T_e} \right) dt, \quad (4.146)$$

with

$$\kappa \equiv \frac{8\sqrt{\pi}g}{3\sqrt{6}} \left(\frac{e^2}{4\pi\epsilon_0} \right)^3 \frac{n_e^2}{m_e T_\gamma^3 \sqrt{m_e T_e}},$$

where $g \sim 2$ is called the Gaunt factor.

The current constraints on these three kinds of distortions [48] are mainly obtained from the analysis of the FIRAS spectrum, and give

$$|y| < 1.9 \times 10^{-5}, \quad |\tilde{\mu}| < 9 \times 10^{-5}, \quad |Y_F| < 1.5 \times 10^{-5}. \quad (4.147)$$

These strongly constrain any processes capable of injecting energy into the Universe from $z \sim 10^6$ onwards, and even from primordial nucleosynthesis (see Fig. 4.13).

4.4.2.3 Monopole and dipole

The observations by FIRAS [49] depend on the position and frequency. They have been expanded in spherical harmonics into monopole, dipole and residual anisotropies.

Even though each point in the sky has a Planck spectrum, the spectrum temperature is slightly higher in one half of the sky and lower in the other half. This dipolar distortion was measured by FIRAS

$$\delta T(\theta) = (3.346 \pm 0.017) \times 10^{-3} \text{ K} \cos \theta,$$

and is compatible with a local kinematic origin associated with the motion of the Solar System and of our Galaxy with respect to the CMB rest frame. The dipole would then arise from a Doppler effect, i.e. from a temperature fluctuation of the form

$$\frac{T}{T_0} = \left[\sqrt{1 - v^2/c^2} \left(1 - \frac{v}{c} \cos \theta \right) \right]^{-1} \sim 1 + \frac{v}{c} \cos \theta + \frac{1}{2} \frac{v^2}{c^2} \cos 2\theta + \mathcal{O}\left(\frac{v^3}{c^3}\right),$$

where θ is the angle between the line of sight and the dipole, i.e. the direction of motion. We thus conclude that the centre of gravity of the Solar System moves in the

direction $(l, b) = (263.85^\circ \pm 0.10^\circ, 48.25^\circ \pm 0.04^\circ)$ in galactic coordinates, with velocity $v \sim 368 \pm 2 \text{ km} \cdot \text{s}^{-1}$. The corresponding velocity of our Galaxy and the local group is then $v \sim 627 \pm 3 \text{ km} \cdot \text{s}^{-1} \sim 0.0022c$ in the direction $(l, b) = (276^\circ \pm 3^\circ, 30^\circ \pm 3^\circ)$.

From this dipole, we can determine the absolute motion of the satellite with respect to the CMB rest frame, i.e. the reference frame for which no dipole is present. Note that this dipole could have a cosmological origin (its effect would not be distinguishable from that of a kinematic dipole). However, since the value of the quadrupole is around 1% of the dipole, the kinematic interpretation is commonly accepted.

4.4.2.4 Residual fluctuations

Residual fluctuations

$$\delta T(\theta, \varphi) = T(\theta, \varphi) - T_0, \quad (4.148)$$

with characteristic amplitude of $\langle (\delta T/T)^2 \rangle^{1/2} \sim 1.1 \times 10^{-5}$, have also been detected by the COBE satellite.

To reach these fluctuations, various foreground effects had to be eliminated first (dust, galactic emission, synchrotron, free-free,...). The spectrum of these emissions is very different from a Planck spectrum. Fig. 4.15 represents their relative amplitudes as a function of the frequency. These contributions can thus be subtracted (more or less easily).

After having subtracted these effects, some temperature anisotropies remain, with relative amplitude $\sim 10^{-5}$, which correspond to temperature fluctuations of the order of $30 \mu\text{K}$. These anisotropies correspond to anisotropies in the cosmic microwave background at the time of recombination and redshifted by the cosmic expansion. Chapter 6 is dedicated to the study of these temperature anisotropies.

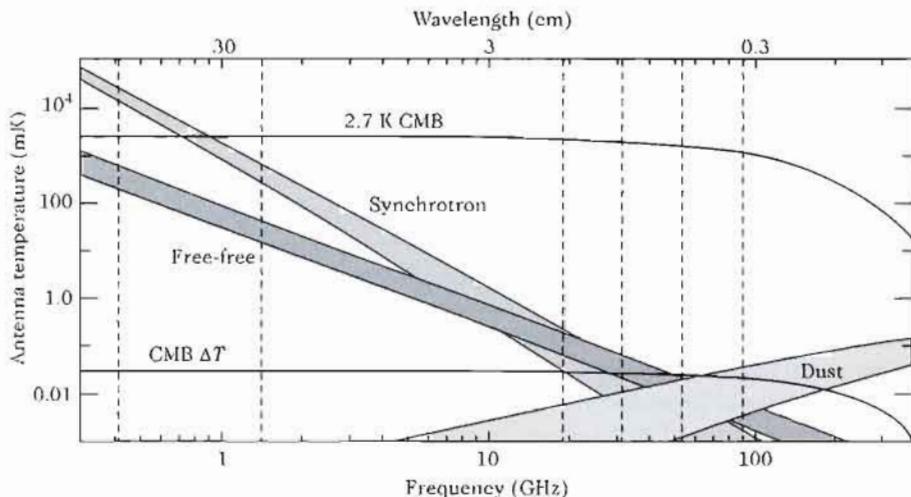


Fig. 4.15 Different foreground effects spoiling the cosmic microwave background. None of these effects have a Planck spectrum. From Ref. [48].

To conclude, the cosmic microwave background confirms that the Universe emerges from a state in thermodynamic equilibrium and that its matter was ionized in the past. Its temperature allows us to determine the photons density. This contribution represents around 93% of the electromagnetic radiation, integrated over all wavelengths. We can thus deduce the radiation density, the best measured cosmological parameter. The cosmic microwave background can be used to strongly constrain any energy injection since $z \sim 10^6$.

4.4.3 Another proof of the expansion of the Universe

The evolution of the temperature of the cosmic microwave background gives an additional proof of the expansion of the Universe. Indeed, it must scale as the inverse of the scale factor, so that

$$T(z) = 2.725(1 + z) \text{ K}. \quad (4.149)$$

These photons can excite the fine structure levels of some atoms. The temperature of the CMB photons at a redshift of $z = 2.33771$ has recently been measured [50] from the observation of absorbing clouds. The measurement is based on the excitation of the two first hyperfine levels of neutral carbon. These excitations are induced by collisions and by the tail of the CMB photon distribution. After subtracting the first contribution, which can be evaluated from other transitions, one can determine the distribution temperature of the relic photons. It leads to the constraint

$$6.0 \text{ K} < T < 14 \text{ K}, \quad (4.150)$$

while the theoretical prediction is $T = 9.1 \text{ K}$. Fig. 4.16 summarizes the various attempts to check the relation (4.149) and proves that it seems to be satisfied. This provides an additional proof of the expansion of the Universe.

4.5 Status of the Big-Bang model

4.5.1 A good standard model...

The Friedman–Lemaître solutions derived in Chapter 3 describe homogeneous and isotropic expanding Universes. This expansion is now established by various observations (see Section 4.1.1) and the Hubble constant has been measured with a good precision. Questioning the reality of this expansion would require the reinterpretation of various observations, and this has not yet been achieved by any alternative model (see, for example, Ref. [51] for an original and instructive attempt). The expansion of the Universe and the redshift represent the first observational pillar of the Big-Bang model.

The cosmological principle, i.e. the homogeneity and isotropic hypothesis, imposes the symmetries of cosmological space-times and is at the origin of their mathematical simplicity. This hypothesis is supported by the observations of the cosmic microwave background that prove the high level of isotropy of the Universe around us. The Big-Bang model does not address the origin of this homogeneity and isotropy. It is thus an important issue to further test the Copernican principle observationally.

The study of nuclear and electromagnetic processes in cosmological space-times has led to two main conclusions: (1) the existence and abundances of light nuclei are

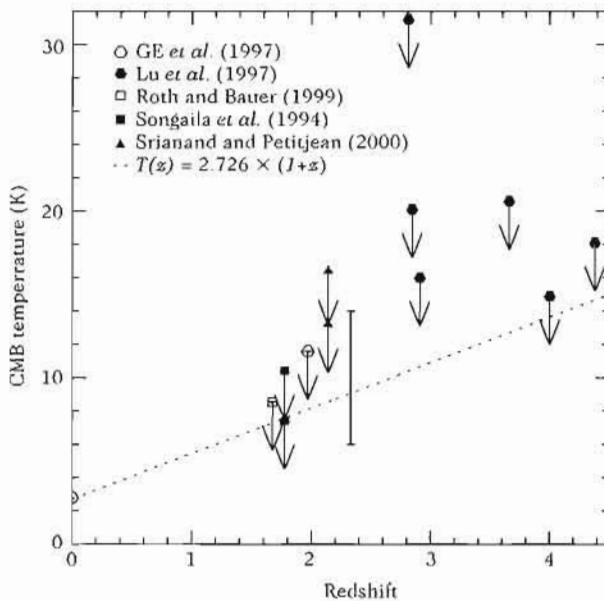


Fig. 4.16 Measurements of the temperature of the cosmic microwave background at different redshifts. The dashed line represents the prediction from the Big-Bang model. From Ref. [50].

naturally explained to within ten orders of magnitude, Section 4.3; and (2) there must exist a cosmic microwave background of photons of cosmological origin, Section 4.4. Both these predictions have been confirmed by observations, providing two other observational pillars of the model.

As already seen, out-of-equilibrium mechanisms play a central role in these predictions. These conclusions rely on sensible, even conservative, hypotheses, namely the validity of general relativity and nuclear and electromagnetic physics at temperatures lower than 100 MeV, providing a robust description of the history of the Universe from 10^{-4} s after the Big Bang.

This Big-Bang model is therefore an excellent *standard model* for the description of the Universe. We will use it as a starting point in the rest of this book and it is nowadays adopted unanimously. A number of conclusions are generic to this model:

- the Universe was dominated by radiation in the past. Thus, a radiation-dominated era and matter-dominated era succeed each other. Then, there can possibly be an era dominated by a cosmological constant or by the curvature of the Universe.
- the Universe has a thermal history. During its expansion, the Universe cools down. The interactions that are efficient at high temperatures tend to decouple as soon as their reaction rate becomes smaller than the expansion rate.
- the Universe emerges from a state where matter at high temperature is ionized and is in thermodynamic equilibrium.

The dynamics of the Big-Bang models depend on their matter content. As seen in

Chapter 3, all background observables depend on the function $E(z)$ only. A detailed assessment of this content can be made [52]. As seen in this chapter, Big-Bang models are characterized by 5 independent parameters ($h, \Omega_K, \Omega_m, \Omega_b, \Omega_r, \Omega_\Lambda$) with the constraint that the sum over the Ω s should be 1. The values of these parameters, as deduced from observations, are summarized in Table 4.7. From these parameters, a few quantities can be deduced that characterize the thermal history of the Universe. They are summarized in Table 4.8.

Table 4.7 Cosmological parameters of Big-Bang models and constraints discussed in this chapter.

| | | | |
|----------------------------|--|--|---------------|
| h | Hubble constant | 0.72 ± 0.05 | Section 4.1.1 |
| Ω_{K0} | curvature parameter | -0.0049 ± 0.013 | Chapter 6 |
| $\Omega_{\mathrm{c}0} h^2$ | dark-matter density | 0.113 ± 0.009 | Chapter 7 |
| $\Omega_{\mathrm{b}0} h^2$ | baryon density (BBN) | 0.018 ± 0.06 | (4.119) |
| | baryon density (CMB) | 0.0224 ± 0.009 | |
| $\Omega_{\gamma0} h^2$ | photon density | $2.4696 \times 10^{-5} \Theta_{2.7}^4$ | (4.141) |
| $\Omega_{\nu0} h^2$ | (massless) neutrino density (3 families) | $1.68 \times 10^{-5} \Theta_{2.7}^4$ | (4.47) |
| $\Omega_{r0} h^2$ | radiation density | $4.148 \times 10^{-5} \Theta_{2.7}^4$ | |
| $\Omega_{\Lambda 0}$ | cosmological constant | 0.72 ± 0.03 | Section 4.1.2 |
| w | dark-energy equation of state | $w < -0.6$ | (4.4) |

Table 4.8 Characteristic quantities of the thermal history of the Universe.

| | | | |
|---------------------------|--------------------------------|---|----------------|
| η | baryon/photon ratio (CMB) | $(6.14 \pm 0.25) \times 10^{-10}$ | (4.120) |
| | baryon/photon ratio (BBN) | $(5.15 \pm 1.75) \times 10^{-10}$ | (4.119, 4.142) |
| T_0 | present CMB temperature | 2.725 ± 0.001 K | (4.137) |
| z_{eq} | redshift at equality | $3612(\Omega_{\mathrm{m}0} h^2 / 0.15) \Theta_{2.7}^{-3}$ | (4.8) |
| T_{eq} | temperature at equality | $5.65 \Theta_{2.7}^{-3} (\Omega_{\mathrm{m}0} h^2)$ eV | (4.8) |
| z_{dec} | redshift at decoupling | 1089 ± 1 | Section 4.4 |
| Δz_{dec} | width | 195 ± 2 | Section 4.4 |
| T_{dec} | temperature at decoupling | 2970 K | Section 4.4 |
| z_{BBN} | redshift at nucleosynthesis | $\sim 10^{10}$ | Section 4.3 |
| T_{BBN} | temperature at nucleosynthesis | ~ 1 MeV | Section 4.3 |

4.5.2 ... but an incomplete model

The Big-Bang model offers a convincing picture of the evolution of the Universe. At higher temperatures the model is less reliable since the laws of physics have to be extrapolated to regimes where they have not been tested otherwise. For example, we still do not have any model explaining the origin of baryons. But it is sensible to hope that we will be able to extrapolate, at least qualitatively, up to the grand unification scale, i.e. to around 10^{16} GeV. Beyond this value, probably at energies close to the

Planck scale at 10^{19} GeV, we expect to have to consider quantum gravity effects. We also get into more and more speculative zones but that gives us the possibility to test the phenomenology of theoretical models at these energies. Putting this aside, the Big-Bang model has a list of problems inherent to its formulation.

4.5.2.1 Flatness problem

The equation for the evolution of the curvature (3.53) is of the form

$$\frac{d\Omega_K}{d \ln a} = (3w + 1)(1 - \Omega_K)\Omega_K, \quad (4.151)$$

if we neglect the recent contribution from the cosmological constant. If w is constant, this equation can be integrated easily

$$\Omega_K = \frac{\Omega_{K0}}{(1 - \Omega_{K0})(1 + z)^{1+3w} + \Omega_{K0}}.$$

So that if the present curvature is small, i.e. $|\Omega_{K0}| = |\Omega_0 - 1| < 0.1$, as observations tend to prove, then we need

$$|\Omega(z_{\text{eq}}) - 1| < 3 \times 10^{-5} \quad (4.152)$$

at equality and

$$|\Omega(z_{\text{pl}}) - 1| < 10^{-60} \quad (4.153)$$

at the Planck time. The Big-Bang model does not give any explanation for such a small curvature at the beginning of the Universe. We could have ended with the same conclusion from the dynamical analysis of Section 3.2.2 that concluded that, for any equation of state $w \geq 0$, the fixed point $\Omega_K = 0$ was unstable.

We should therefore determine a mechanism leading to such a small curvature or explain why we strictly have $\Omega_K = 0$. One solution is already suggested in Fig. 3.6 that illustrates that $\Omega_K = 0$ is an attractor if the Universe is dominated by a matter component with equation of state $w < -1/3$.

In the primordial era, as long as the Universe is dominated by ordinary matter, it is always very flat, so that the flatness problem can actually be seen as an age problem: why does our Universe look so young?

4.5.2.2 Horizon problem

The cosmological principle, which imposes space to be homogeneous and isotropic, is at the heart of the Friedmann and Lemaître solutions. By construction, they cannot explain the origin of this homogeneity and isotropy. So it would be more satisfying to find a justification of this principle.

From an observational point of view, the CMB radiation supports this hypothesis. Its isotropy could be easily explained if the photons were emitted once the causal contact was established between them. For that, the photons should be in causal contact since the time of decoupling. To compute the size of the causally connected regions at decoupling, we work in conformal time (see Fig. 3.12).

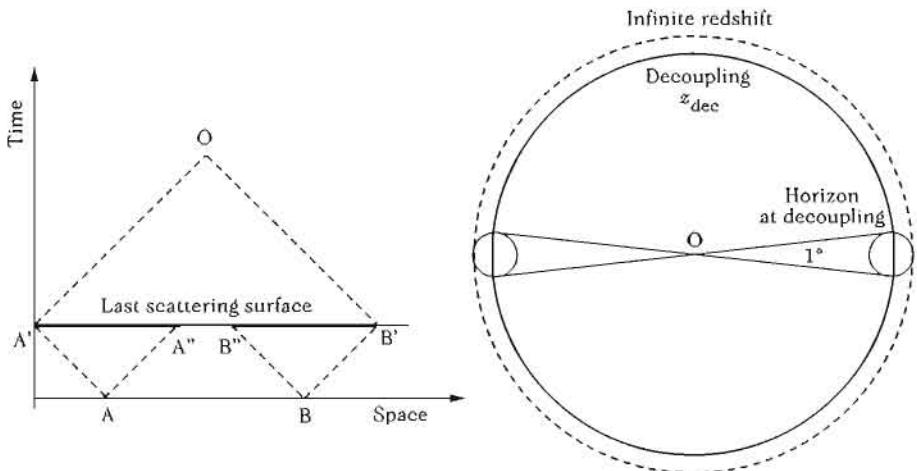


Fig. 4.17 (left): The past light cone of an observer O intersects the surface of last scattering on a sphere of diameter $A'B'$ that determines the size of the observable Universe. The size of the causally connected regions ($A'A''$, ...) is determined by the intersection with the future light cone from the Big-Bang hypersurface. (right): The horizon at decoupling and the surface of last scattering.

As can be seen from Fig. 4.17, the regions in causal contact at the time of decoupling are smaller than the size of the observable Universe. Thus, the surface of last scattering includes around

$$N \sim \left(\frac{\eta_0 - \eta_{\text{dec}}}{\eta_{\text{dec}}} \right)^3 \sim 8(1 + z_{\text{dec}})^{3/2} \sim 10^5 - 10^6 \gg 1$$

regions in causal contact. Such a causal region at decoupling is now observed under an angle

$$\theta \sim 2 \frac{\eta_{\text{dec}}}{\eta_0 - \eta_{\text{dec}}} \sim 1^\circ. \quad (4.154)$$

It is difficult to explain why the temperature of the cosmic microwave background is the same up to $10^{-3}\%$ in the entire sky, while the latter is composed of about 10^6 causally independent regions. The standard cosmological model predicts that only small regions of the surface of last scattering should be correlated, whereas larger scales should necessarily be uncorrelated.

Another way to formulate the horizon problem is to give an estimate of the number of initial cells, with an initial characteristic Planck size length, present today in the observable Universe. This number is typically of order

$$N \sim \left(\frac{1 + z_{\text{pl}}}{\ell_{\text{pl}} H_0} \right)^3 \sim 10^{87}.$$

Here again, the study of galaxies tends to show that their distribution is homogeneous on larger scales, so it is difficult to understand how initial conditions fixed on 10^{87} causally independent regions can appear so identical (at a 10^{-5} level!).

The horizon problem is related to the state of thermodynamic equilibrium in which the Universe is, discussed in Section 4.2. The cosmological principle imposes a non-causal initial condition on spatial sections of the Universe and in particular that the temperature of the thermal bath is the same at every point.

The horizon problem is thus closely related to the cosmological principle and is therefore deeply rooted in the Friedmann–Lemaître solutions.

4.5.2.3 Problem of the origin of structures

Friedmann–Lemaître solutions describe strictly homogeneous and isotropic space-times. However, the Universe is filled with numerous structures, whose origin and properties cannot be addressed and explained within this framework. As formulated, the cosmological model is thus incomplete.

Since gravity is purely attractive, any density perturbation will tend to collapse to create structure. We can thus think that small fluctuations, of thermal origin, for example, could explain the large-scale structure. But this mechanism is actually not very efficient in an expanding Universe (see Chapter 5). As long as the Universe is radiation dominated, radiation pressure acts against gravitational collapse, and hence delays the beginning of large-scale structure formation. Thus, density fluctuations can only grow significantly for $z > z_{\text{eq}} \sim 10^4$. We should therefore find a mechanism able to generate density fluctuations with non-negligible amplitude. Such fluctuations do indeed exist and have been detected by the COBE satellite.

Thus, a new horizon problem appears. In a radiation- and matter-dominated Universe, the causal radius, roughly the Hubble radius, increases more rapidly than the physical wavelength of perturbations. Thus, the wavelengths of all cosmological perturbations observed today must have been larger than the causal radius at the beginning of the expansion of the Universe. In particular, the satellite COBE has detected temperature fluctuations with wavelengths larger than the horizon at decoupling. Moreover, these fluctuations seem to have a scale-invariant spectrum.

Can these initial density perturbations be generated by a causal mechanism? How can we explain their amplitude and their spectrum?

Another problem related to the existence of these perturbations is the *trans-Planckian problem*. All physical wavelengths observed today were smaller than the Planck scale at some time in the past. Thus, they must have been described by some physics that we do not know. Until which point are cosmological observations robust enough against the hypothesis of this (unknown) physics? Can they be used to extract some information on this physics? Note that the Hubble radius also becomes smaller than the Planck length, it is thus unclear whether the description of the Universe by a classical space-time is still acceptable. This problem is related to the existence of an initial singularity. This version of the trans-Planckian problem is slightly different from the one presented in Chapter 8 where only the wavelength of perturbations becomes trans-Planckian, whereas the Hubble radius remains larger than the Planck length.

4.5.2.4 Relic problem

While the Universe cools down, some symmetries can be spontaneously broken (see Chapter 11). During these phase transitions, topological defects can be formed.

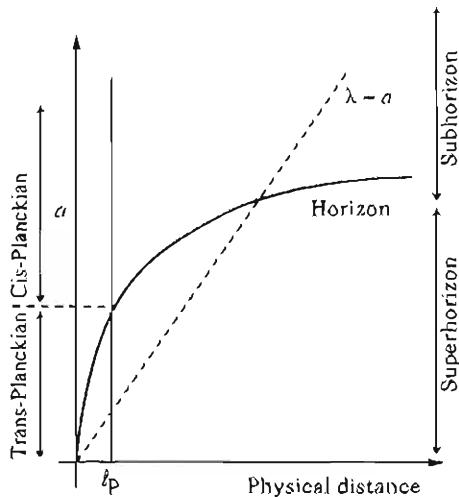


Fig. 4.18 Horizon problem for a perturbation with physical wavelength λ and the trans-Planckian problem related to the existence of a singularity. All physical lengths, in particular wavelengths, become smaller than the Planck scale. The Hubble radius also becomes smaller than the Planck scale and the classical description of space-time can be questioned in this limit.

In particular, in the grand unified theory framework, non-gravitational interactions are assumed to have a symmetry described by a group \mathcal{G} . This group must be broken to give rise to the one observed at low-energy, namely $SU(3)_c \times SU(2)_L \times U(1)_Y$ (see Chapter 2). If the group \mathcal{G} is a simple group, then during this symmetry breaking, monopoles are produced with a mass of order $M_{\text{monopole}} \sim T_{\text{GUT}} \sim 10^{16} \text{ GeV}$. Since the fields giving rise to these monopoles are not a priori correlated on distances larger than the causal horizon, we can assume that they are created with a unit density per volume during the phase transition, $n_{\text{monopole}} \sim (2t_{\text{GUT}})^{-3}$. We get that the monopole density is $\rho_{\text{monopole}} \sim M_{\text{monopole}} n_{\text{monopole}}$.

At t_{GUT} , the Universe is dominated by radiation, so that $t_{\text{GUT}} \sim T_{\text{GUT}}^{-2}$ (see Chapter 11 for a description of phase transitions in the cosmological framework). It follows that

$$\Omega_{\text{monopole}} h^2 \sim 10^{17} \left(\frac{T_{\text{GUT}}}{10^{16} \text{ GeV}} \right)^3 \left(\frac{M_{\text{monopole}}}{10^{16} \text{ GeV}} \right). \quad (4.155)$$

As soon as they are formed, such relics would therefore overdominate the matter content of the Universe. The effects of this matter on the evolution of the Universe would be catastrophic. This problem was initially referred to as the *monopole problem*.

To conclude, the overproduction of massive relics in the primordial Universe is expected in various theoretical frameworks, the grand unified theory in particular. We should explain why these relics do not affect the dynamics of the Universe.

4.5.2.5 Dark-sector problem

According to the conclusions from cosmological observations, only around 5% of the cosmic matter, so-called baryonic, is in a form expected by the standard model of particle physics. We should postulate the existence of a dark sector representing 95% of the matter and having at least two components:

- the dark matter, representing $23 \pm 3\%$ of the matter. This non-relativistic matter cannot be in baryonic form. The production of relics in the primordial Universe provides a mechanism that could lead to such matter. We should determine the interactions involved and the nature of this particle.
- the dark energy, representing $72 \pm 3\%$ of matter, is required to explain the recent acceleration of the Universe. The cosmological constant has been a natural candidate for a long time.

A satisfying cosmological model cannot fail to have a deep understanding of this dark sector. In particular, we should understand why dark matter and baryonic matter have comparable abundances. We should also remember that baryogenesis is still not a well-understood process. This mechanism should explain why $\eta = n_b/n_\gamma \sim 10^{-10}$.

As for the dark energy, two possibilities are to be considered:

- the cosmological constant is a manifestation of the vacuum. In this case, its energy density can be computed from the theory of particle physics. However, any vacuum energy coming from a theory compatible with the standard model of particle physics should be at least of order of the energy of the electroweak symmetry-breaking scale, $\rho_\Lambda^{1/4} \sim 1$ TeV. But the cosmological value is of the order of $\rho_\Lambda^{1/4} \sim 10^{-3}$ eV. The disagreement between the theoretical estimate and the cosmologically determined value is thus at least of the order of 10^{60} , and can go up to 10^{120} if we extrapolate to the Planck scale. This problem is known as the *cosmological constant problem*.
- The dark energy is a dynamical quantity. These models will be described in detail in Chapter 13. However, we could ask ourselves why this matter only starts to dominate the matter content of the Universe today. This is what is called the *coincidence problem*.

The coincidence problem is actually more general and relies on the fact that the different matter components are in the following ratio today

$$\Omega_{\gamma 0} : \Omega_{b0} : \Omega_{c0} : \Omega_{\Lambda 0} \sim 10^{-3} : 1 : 5 : 14. \quad (4.156)$$

How can we understand that these components of matter, whose origin depends a priori on differing mechanisms, come with such a ratio? Why do we live at a time when the two components of the dark sector are in comparable quantities? Also, note that the existence of structure and the cosmological observables are very sensitive to these ratios.

To finish, note that any information obtained on this dark sector has been deduced from the observation of luminous matter. These conclusions rely on a central hypothesis: the validity of general relativity to describe gravity. This hypothesis is so central that it is mandatory to test it on cosmological scales. This last point merges with

the effective constancy of the fundamental constants of Nature on cosmological time scales.

Note once more that these conclusions have been obtained by imposing that the Universe is well described by a Friedmann-Lemaître space. Most observations, and in particular the supernovæ data, from which we deduce the recent acceleration of the Universe, are localized in our past light cone. This implies that dependencies in space and time are in general degenerate. This degeneracy is lifted by the symmetry hypothesis in the framework of Friedmann-Lemaître models. The conclusions could thus be different in a Universe without these symmetries. For instance, the supernovæ data can be reproduced in a Universe with spherical symmetry, filled only with dust (Lemaître-Tolman-Bondi space described in Chapter 3). The expansion of such a Universe is not accelerated but the observation of supernovæ would be interpreted with a wrong symmetry hypothesis and would lead to the conclusion that the expansion is accelerating, whereas it actually only reflects the spatial dependance of the metric. Once more, we should check the Copernican principle as much as possible and verify all conclusions that can be deduced from the observations: they depend on the hypothesis made on the space-time structure.

4.5.3 Conclusion

The Big-Bang model is a good standard model, giving a satisfactory description of the Universe on large scales, where it can be described by a homogeneous and isotropic space-time. It has, nonetheless, various problems that primordial cosmology aims to resolve. At higher temperatures, physics is less well understood, and its cosmological implications can offer solutions to some of these problems, but can also constrain this physics.

It is thus important for primordial cosmology to have a standard model such as the Big Bang. This starting point will allow us to formulate some variations, to test them and to study their phenomenology. The rest of this book is dedicated to this program.

References

- [1] R.A. ALPHER, H. BETHE and G. GAMOW, 'The origin of chemical elements', *Phys. Rev. Lett.* **73**, 803, 1948.
- [2] G. GAMOW, 'The evolution of the Universe', *Nature* **162**, 680, 1948.
- [3] R.A. ALPHER and R. HERMAN, 'Evolution of the Universe', *Nature* **162**, 774, 1948.
- [4] E. HUBBLE, 'A relation between distance and radial velocity among extra-galactic nebulae', *Proc. Natl. Acad. Sci. USA* **15**, 168, 1929.
- [5] S.K. WEBB, *Measuring the Universe, the cosmological distance ladder*, Springer-Verlag, 1999.
- [6] W. FREEMAN, 'The Hubble constant and the expansion age of the Universe', *Phys. Rep.* **333**, 13, 2000.
- [7] W. FREEMAN *et al.*, 'Final results from the Hubble telescope key project to measure the Hubble constant', *Astrophys. J.* **553**, 47, 2001.
- [8] M. BIRKINSHAW, 'The Sunyaev-Zeldovich effect', *Phys. Rep.* **310**, 97, 1999.
- [9] S. PERLMUTTER *et al.*, 'Measurement of Ω and Λ from 42 high- z supernovæ', *Astrophys. J.* **517**, 565, 1999.
- [10] A.G. RIESS *et al.*, 'Observational evidence from supernovæ for an accelerating Universe and a cosmological constant', *Astron. J.* **116**, 1009, 1998.
- [11] A. FILIPPENKO, 'The accelerating Universe and dark energy: evidence from type Ia supernovæ', *Lect. Notes Phys.* **646**, 191, 2004.
- [12] U. SELJAK *et al.*, 'Cosmological parameter analysis including SDSS Ly- α forest and galaxy bias: constraints on the primordial spectrum of fluctuations, neutrino mass, and dark energy', *Phys. Rev. D* **71**, 103515, 2005.
- [13] B. CHABOYER, 'The age of the Universe', *Phys. Rep.* **307**, 23, 1998.
- [14] J.J. COWAN, F.-K. THIELEMANN and J.W. TRURAN, 'Radioactive dating of the elements', *Annu. Rev. Astron. Astrophys.* **29**, 447, 1991.
- [15] R. CAYREL *et al.*, 'Measurements of stellar ages from uranium decay', *Nature* **409**, 691, 2001.
- [16] E. KOLB and M. TURNER, *The early Universe*, Addison Wesley, 1990.
- [17] R. TOLMAN, *Relativity, thermodynamics and cosmology*, Cambridge University Press, 1934.
- [18] J. BERNSTEIN, *Kinetic theory in the expanding Universe*, Cambridge University Press, 1988.
- [19] P.J.E. PEEBLES, 'Primordial helium abundance and the primeval fireball', *Astrophys. J.* **146**, 542, 1966.
- [20] J. BERNSTEIN, L.S. BROWN and G. FEINBERG, 'Cosmological helium production simplified', *Rev. Mod. Phys.* **61**, 25, 1989.
- [21] P.J.E. PEEBLES, *Physical cosmology*, Princeton University Press, 1971.

- [22] S. WEINBERG, *Gravitation and cosmology: principles and applications of the general theory of relativity*, John Wiley and Sons, 1972.
- [23] S. SARKAR, 'Big-Bang nucleosynthesis and physics beyond the standard model', *Rep. Prog. Phys.* **59**, 1493, 1996.
- [24] K.A. OLIVE, G. STEIGMAN and T.P. WALKER, 'Primordial nucleosynthesis: theory and observations', *Phys. Rep.* **333**, 389, 2000.
- [25] V. MUKHANOV, 'Nucleosynthesis without a computer', *Int. J. Theor. Phys.* **43**, 669, 2004.
- [26] R.V. WAGONER, 'On the synthesis of elements at very high temperatures', *Astrophys. J.* **148**, 3, 1967.
- [27] <http://www-thpbphys.physics.ox.ac.uk/users/SubirSarkar/bbn.html>.
- [28] K.M. NOLLETH and S. BURLES, 'Estimating reaction rates and uncertainties from primordial nucleosynthesis', *Phys. Rev. D* **61**, 123505, 2000.
- [29] Y.I. IZOTOV et al., 'Helium abundance in the most metal deficient blue compact galaxies', *Astrophys. J.* **527**, 757, 1999.
- [30] V. LURIDJANA et al., 'The effect of collisional enhancement of Balmer lines on the determination of the primordial helium abundance', *Astrophys. J.* **592**, 846, 2003.
- [31] Y.I. IZOTOV and T.X. THUAN, 'Systematic effects and a new determination of the primordial abundance of ^4He and D/Z from observations of blue compact galaxies', *Astrophys. J.* **602**, 200, 2004.
- [32] D. KIRKMAN et al., 'The cosmological baryon density from the deuterium to hydrogen towards QSO absorption system: D/H towards Q1243+3047', *Astrophys. J. Suppl.* **149**, 1, 2003.
- [33] S.G. RYAN et al., 'Primordial lithium and big Bang nucleosynthesis', *Astrophys. J.* **523**, 654, 1999.
- [34] D.N. SPERGEL et al., 'First year Wilkinson microwave anisotropy probe (WMAP) observations: determination of cosmological parameters', *Astrophys. J. Suppl.* **148**, 175, 2003.
- [35] A. COC et al., 'Updated big Bang nucleosynthesis confronted to WMAP observations and to the abundance of light elements', *Astrophys. J.* **600**, 544, 2004.
- [36] K. OLIVE and E. SKILLMAN, 'A realistic determination of the error of primordial helium abundance: steps toward non-parametric nebular helium abundances', *Astrophys. J.* **617**, 29, 2004.
- [37] J. LINSKY, 'Atomic deuterium/Hydrogen in the Galaxy', *Space Sci. Rev.* **106**, 49, 2003; H.W. MOOS et al., 'Abundances of deuterium, oxygen, and nitrogen in the local interstellar medium: overview of first results from FUSE mission', *Astrophys. J. Suppl.* **140**, 3, 2002.
- [38] T.M. BANIA, R.T. ROOD and D.S. BALSER, 'The cosmological density of baryons from observations of $^3\text{He}^+$ in the Milky Way', *Nature* **415**, 54, 2002.
- [39] R. CYBURT et al., 'New BBN limits on physics beyond the standard model from He^4 ', *Astropart. Phys.* **23**, 313, 2005.
- [40] A. PENZIAS and R. WILSON, 'A Measurement of excess antenna temperature at 4080 Mc/s', *Astrophys. J.* **142**, L419, 1965.

- [41] R. DICKE, P. PEEBLES, P. ROLL and D. WILKINSON, 'Cosmic blackbody radiation', *Astrophys. J.* **142**, 383, 1965.
- [42] G. SMOOT, *The CMB spectrum*, in *The cosmic Microwave Background*, C.H. Lineweaver et al.(eds.), NATO ASI series 502, Kluwer Academic Publishers, 1997.
- [43] F.R. BOUCHET, J.-L. PUGET and J.-M. LAMARRE, *The cosmic microwave background*, in *L'Univers primordial*, Springer, 1999, pp. 103.
- [44] R.B. PARTRIDGE, *3K: the cosmic microwave background*, Cambridge University Press, 1999.
- [45] P.J.E. PEEBLES, 'Recombination of the primeval plasma', *Astrophys. J.* **153**, 1, 1968.
- [46] B.J. JONES and R.F. WISE, 'Ionization of the primeval plasma', *Astron. Astrophys.* **149**, 144, 1985.
- [47] J.C. MATHER, 'Calibrator design of the COBE far infrared absolute spectrophotometer', *Astrophys. J.* **512**, 511, 1999.
- [48] G.F. SMOOT and D. SCOTT, 'Cosmic background radiation', in *The review of particle physics*, S. Eidelman et al., *Phys. Lett. B* **592**, 1, 2004.
- [49] D. FIXSEN et al., 'The cosmic microwave background spectrum from full COBE FIRAS data set', *Astrophys. J.* **473**, 576, 1996.
- [50] R. SRIANAND, P. PETITJEAN and C. LEDOUX, 'The microwave background temperature at the redshift of 2.3371', *Nature* **408**, 931, 2000.
- [51] G.F.R. ELLIS, R. MAARTENS and S.D. NEL, 'The expansion of the Universe', *Month. Not. R. Astron. Soc.* **184**, 439, 1978.
- [52] M. FUKUGITA and P.J.E. PEEBLES, 'The cosmic energy inventory', *Astrophys. J.* **616**, 643, 2004.

5

The inhomogeneous Universe: perturbations and their evolution

Friedmann–Lemaître space-times are approximations that give a good description of our Universe only on very large scales. In fact, our Universe is not strictly homogeneous and isotropic since it contains structures such as galaxies, galaxy clusters, etc. In the 1980s, it was understood that the large-scale structure can result from the gravitational amplification of the density fluctuations of a pressureless fluid, the cold dark matter, see Refs. [1–3].

The aim of this chapter is to study Universes with a geometry ‘close’ to a Friedmann–Lemaître space-time and to understand the evolution of the density perturbations under the influence of gravity in an expanding space-time. To do so, we will start by recalling the Newtonian limit in Section 5.1. Then, we will focus on the general relativistic case in Section 5.2 to present the theory of cosmological perturbations. This part, which is quite technical, is one of the cornerstones of modern cosmology, and is essential in order to address the properties of the large-scale structure of the Universe. In Section 5.3, we will study some solutions of these perturbation equations. The characteristic scales that should be taken into account in the general case, and the form of the power spectrum are described in Section 5.4. We will finish in Section 5.5 with a description of the observations of these large-scale structures.

References [4–9] can complement this chapter.

5.1 Newtonian perturbations

Before developing the general case of the evolution of perturbations in the relativistic framework, let us focus on the simplified case of the evolution of perturbations of an initially homogeneous fluid, in the Newtonian regime.

In fact, as we will see in Section 5.4, such a Newtonian approach is sufficient to describe the late time evolution of the density perturbations during the matter era on sub-Hubble scales. A general description of gravitational dynamics in the Newtonian framework can be found in Ref. [4].

5.1.1 Case of a static space

In a static Euclidean space, the hydrodynamic equations take the usual form [10] of a continuity equation and the Euler equation

$$\partial_t \rho + \nabla \cdot (\rho \mathbf{v}) = 0, \quad (5.1)$$

$$\partial_t \mathbf{v} + (\mathbf{v} \cdot \nabla) \mathbf{v} = -\frac{1}{\rho} \nabla P - \nabla \Phi, \quad (5.2)$$

for a fluid with density ρ , pressure P and velocity field \mathbf{v} , evolving in a gravitational field with potential Φ .

5.1.1.1 Without gravity

If we neglect gravity and decompose the density in terms of a homogeneous component and a perturbation as $\rho = \bar{\rho} + \delta\rho$ and the pressure similarly as $P = \bar{P} + \delta P$, (5.1) and (5.2) reduce to a unique wave equation

$$\partial_t^2 \delta\rho - \Delta \delta P = 0. \quad (5.3)$$

Let us introduce $\chi_a \equiv (\partial P / \partial \rho)_a / \rho$, called the isothermal or adiabatic compressibility coefficient, where a stands either for the temperature (T) or the entropy (S). When the propagation time of heat is brief, the flow can be considered to be isothermal, whereas it will be adiabatic if there was insufficient time to reach equilibrium, which is the case for long wavelengths. In the latter case, we infer that the wave equation (5.3) takes the form

$$\boxed{\partial_t^2 \delta\rho - c_s^2 \Delta \delta\rho = 0, \quad c_s^2 = \rho \chi_S = \left(\frac{\delta P}{\delta \rho} \right)_S.} \quad (5.4)$$

Thus, hydrodynamic perturbations propagate with constant amplitude and with velocity c_s , the speed of sound of the given fluid.

5.1.1.2 Effect of gravity

Let us now turn on gravity. The gravitational potential is determined by the Poisson equation

$$\Delta \Phi = 4\pi G_N \delta\rho, \quad (5.5)$$

so that the evolution equation for the density fluctuations now has an extra term and takes the form

$$\partial_t^2 \delta\rho - c_s^2 \Delta \delta\rho = 4\pi G_N \bar{\rho} \delta\rho. \quad (5.6)$$

By decomposing the perturbations into plane waves as $\delta\rho \propto \exp[i(\omega t - \mathbf{k} \cdot \mathbf{x})]$, we obtain the dispersion relation

$$\boxed{\omega^2 = \frac{4\pi^2 c_s^2}{\lambda_J^2} \left(\frac{\lambda_J^2}{\lambda^2} - 1 \right), \quad \lambda_J \equiv c_s \sqrt{\frac{\pi}{G_N \bar{\rho}}}.} \quad (5.7)$$

λ_J is the Jeans length above which any perturbation is unstable and grows exponentially as

$$\delta\rho \propto \exp \left[\sqrt{4\pi G_N \bar{\rho} \left(1 - \frac{\lambda_J^2}{\lambda^2} \right)} t \right].$$

Qualitatively, the Jeans length characterizes the transition from a regime for which the evolution of the perturbations is dominated, at long wavelengths, by gravity ($\lambda > \lambda_J$), to a region ($\lambda < \lambda_J$) where gravity is negligible and in which we recover sound waves.

The value of the Jeans length can actually be deduced by considering the two dynamical times of the problem. The time associated with sound waves (i.e. the time taken by this wave to propagate along a distance equal to its wavelength λ) is given by $t_{\text{sound}} \sim \lambda/c_s$. The gravitational collapse time taken by a structure with density $\bar{\rho}$ to collapse is $t_{\text{grav}} \sim (G_N \bar{\rho})^{-1/2}$. Hence, at small wavelengths, one has $t_{\text{sound}} \ll t_{\text{grav}}$ and the pressure has time to compensate the gravitational collapse of the fluid, which thus remains quasi-homogeneous. When λ becomes larger than the Jeans length, t_{sound} is larger than t_{grav} , the pressure can no longer balance gravity and large inhomogeneities can develop.

5.1.1.3 Energetic aspects

From an energetic point of view, each point of the fluid can be seen as a harmonic oscillator whose displacement from its equilibrium position is denoted by s .

The velocity of each element of the fluid is thus \dot{s} and, using the conservation (5.1), the density perturbation is $\delta\rho = -\nabla \cdot s$. The equation of propagation (5.7) can then be integrated to give, after averaging over a volume V , the equation for conservation of energy

$$\langle E_M \rangle = \langle E_K \rangle + \langle E_E \rangle + \langle E_G \rangle, \quad (5.8)$$

where $\langle E_K \rangle = \frac{1}{2} \langle \dot{s}^2 \rangle$ is the mean kinetic energy, $\langle E_E \rangle = \frac{1}{2} c_s^2 \langle \nabla s^2 \rangle$, the mean elastic energy, and $\langle E_G \rangle = -2\pi G_N \bar{\rho} \langle s^2 \rangle$, the mean gravitational energy. The kinetic energy takes the form

$$\frac{1}{2} \langle \dot{s}^2 \rangle = \frac{1}{2} \omega^2 \langle s^2 \rangle, \quad \omega^2 = c_s^2 (k^2 - k_J^2), \quad k_J \equiv \frac{2\pi}{\lambda_J}.$$

We deduce that the total energy per unit volume is

$$\langle E_M \rangle = \frac{1}{2} \langle s^2 \rangle [\omega^2 + c_s^2 (k^2 - k_J^2)].$$

The dispersion relation teaches us that there is equipartition of the energy between the kinetic energy and the potential energy (which has two contributions). The mean energy is thus given by

$$\langle E_M \rangle = \langle s^2 \rangle c_s^2 (k^2 - k_J^2), \quad (5.9)$$

and is, as expected, negative for wavelengths larger than the Jeans length.

5.1.2 Case of an expanding space

In an expanding space, one can introduce the comoving coordinates x by

$$r(t) = a(t)x, \quad (5.10)$$

where $a(t)$ is the scale factor. The velocity field is then given by

$$\mathbf{v}(t) = \dot{r} = Hr + \mathbf{u}, \quad (5.11)$$

where $\mathbf{u} = a\dot{x}$ is the proper velocity.

5.1.2.1 Eulerian formulation

Equations (5.1) and (5.2) can be rewritten by noticing that, in the previous case of a static space-time, the time and space derivative operators defined from t and \mathbf{r} were considered to be independent. This is no longer the case here, and it is convenient to use space derivatives defined with respect to the comoving coordinates \mathbf{x} . Consequently, the gradient $\nabla_{\mathbf{r}}$ is replaced by $\nabla_{\mathbf{x}}/a$, and the time derivative $\partial_t \rho(\mathbf{r}, t)$, that previously assumed constant \mathbf{r} , becomes $\partial_t \rho(\mathbf{x}, t) - H\mathbf{x} \cdot \nabla_{\mathbf{x}}\rho(\mathbf{x}, t)$. The two equations are thus re-expressed in comoving coordinates as

$$\dot{\rho}(\mathbf{x}, t) + 3H\rho(\mathbf{x}, t) + \frac{1}{a}\nabla_{\mathbf{x}}[\rho(\mathbf{x}, t)\mathbf{u}(\mathbf{x}, t)] = 0, \quad (5.12)$$

and

$$\dot{\mathbf{u}}(\mathbf{x}, t) + Hu(\mathbf{x}, t) + \frac{1}{a}(\mathbf{u} \cdot \nabla_{\mathbf{x}})\mathbf{u}(\mathbf{x}, t) = -\frac{1}{a}\nabla_{\mathbf{x}}\Phi(\mathbf{x}, t) - \frac{1}{a\rho}\nabla_{\mathbf{x}}P(\mathbf{x}, t), \quad (5.13)$$

where we have used that $\nabla_{\mathbf{x}} \cdot \mathbf{x} = 3$. If the density contrast, δ , is defined by

$$\rho = \bar{\rho}(t)[1 + \delta(\mathbf{x}, t)], \quad (5.14)$$

then the first equation can be rewritten as the continuity equation

$$\dot{\delta} + \frac{1}{a}\nabla[(1 + \delta)\mathbf{u}] = 0, \quad (5.15)$$

where from now on, unless otherwise stated, we omit the reference to comoving coordinates that is implicit in what follows. After some manipulations, an evolution equation for δ can be deduced from these equations

$$\ddot{\delta} + 2H\dot{\delta} = \frac{1}{\rho a^2}\Delta P + \frac{1}{a^2}\nabla \cdot [(1 + \delta)\nabla\Phi] + \frac{1}{a^2}\partial_i\partial_j[(1 + \delta)u^i u^j]. \quad (5.16)$$

Note that this equation does not make any assumptions on δ and in particular it does not assume δ to be small compared to 1. The gravitational potential is determined by the Poisson equation, which then takes the form

$$\Delta\Phi = 4\pi G_N \bar{\rho} a^2(t) \delta. \quad (5.17)$$

Solving (5.17), one gets that the gravitational potential is given by

$$\Phi(\mathbf{x}, t) = -G_N \bar{\rho} a^2 \int d^3x' \frac{\delta(\mathbf{x}', t)}{|\mathbf{x} - \mathbf{x}'|}. \quad (5.18)$$

These equations provide a Eulerian description of the dynamics.

5.1.2.2 Lagrangian description

A Lagrangian approach can also be developed by following the trajectories of particles or fluid elements instead of considering the dynamics of the density and velocity fields.

The central quantity is then the displacement field $\mathbf{X}(q, t)$, which maps the initial position of a particle labelled q into its Eulerian position at time t according to

$$\mathbf{x}(t) = \mathbf{q} + \mathbf{X}(\mathbf{q}, t).$$

The equation of motion for particle trajectories $\mathbf{x}(t)$ is then

$$\ddot{\mathbf{x}} + 2H\dot{\mathbf{x}} = -\frac{1}{a^2}\nabla_{\mathbf{x}}\Phi,$$

or equivalently in terms of the conformal time η

$$\mathbf{x}'' + \mathcal{H}\mathbf{x}' = -\nabla_{\mathbf{x}}\Phi.$$

Taking the divergence of this equation, one obtains an equation for the displacement field

$$J(q, t)\nabla_{\mathbf{x}} \cdot [\mathbf{X}'' + \mathcal{H}\mathbf{X}'] = \frac{3}{2}\Omega_m\mathcal{H}^2[J(q, t) - 1], \quad (5.19)$$

where we have made use of the Poisson equation and of the fact that $\bar{\rho}d^3q = \bar{\rho}[1 + \delta(\mathbf{x}, t)]d^3x$, which implies that

$$1 + \delta(\mathbf{x}, t) = \frac{1}{\det(\delta_{ij} + X_{i,j})} \equiv \frac{1}{J(q, t)},$$

where $X_{i,j} \equiv \partial X_i / \partial q^j$, and $J(q, t)$ is the Jacobian of the transformation from Eulerian to Lagrangian space. Note that the gradient in Eulerian coordinates can be transformed to a gradient in Lagrangian coordinates as

$$(\nabla_{\mathbf{x}})_i = (\delta_{ij} + X_{i,j})^{-1}(\nabla_{\mathbf{q}})_j.$$

Such a description can be useful for some problems and more details can be found in Refs. [11, 12]. Note, however, that when there is *shell crossing*, that is when the fluid elements with different initial positions q end up at the same Eulerian position x , the Jacobian vanishes and the density field becomes singular. At these points, the description of the dynamics in terms of a mapping no longer holds.

5.1.2.3 Linearized systems

Let us now suppose that the fluid density is only weakly perturbed with respect to the homogeneous configuration. This translates into the two conditions

$$\delta \ll 1, \quad \left(\frac{ut}{d}\right)^2 \ll \delta,$$

where u is the characteristic fluid velocity and d the coherence length of density perturbations. The second condition implies that the gradients must be small. Let us recall that in a flat Universe with no cosmological constant, the Friedmann equation implies that $6\pi G_N \bar{\rho}t^2 = 1$ during the matter-dominated era (since $a \propto t^{2/3}$).

By linearizing (5.16), we get the propagation equation

$$\ddot{\delta} + 2H\dot{\delta} = \frac{c_s^2}{a^2} \Delta \delta + 4\pi G_N \bar{\rho} \delta. \quad (5.20)$$

We recover the Jeans length defined by (5.7). Unlike the case of a static Universe, the Jeans length now depends on time via $\bar{\rho}$. A perturbation with given wavelength will thus be able to pass from a sound-wave regime to a gravitational collapse regime, and this at a time depending on the value of the wavelength.

5.1.2.4 Growth of density perturbations

For a fluid of matter with negligible pressure, (5.20) is a second order linear differential equation involving only time. One can therefore look for solutions of the form

$$\delta(x, t) = D_+(t)\varepsilon_+(x) + D_-(t)\varepsilon_-(x),$$

where the functions $\varepsilon(x)$ correspond to the initial density field. We will refer to D_+ as the growing mode and to D_- as the decaying mode. The functions D are solutions of the equation

$$\ddot{D} + 2H(t)\dot{D} - \frac{3}{2}H^2(t)\Omega_m(t)D = 0, \quad (5.21)$$

where we have expressed the mean matter density as $4\pi G_N \rho_m(t) = \frac{3}{2}H^2\Omega_m(t)$. For a flat Universe with no cosmological constant ($\Omega_m = \Omega_{m0} = 1$, $a \propto t^{2/3}$), one can check that

$$D_+(t) \propto t^{2/3} \propto a(t), \quad D_-(t) \propto t^{-1} \propto a^{-3/2}(t). \quad (5.22)$$

In the general case, this equation needs be integrated numerically. A convenient way to solve it is to use the scale factor, a , as the time coordinate, so that (5.21) takes the form

$$\frac{d^2D}{da^2} + \left(\frac{1}{H} \frac{dH}{da} + \frac{3}{a} \right) \frac{dD}{da} - \frac{3}{2} \frac{\Omega_{m0}}{a^5} \left(\frac{H}{H_0} \right)^{-2} D = 0, \quad (5.23)$$

where we have used the fact that $4\pi G_N \bar{\rho} = 4\pi G_N \rho_{m0} a^{-3}$ and assumed that $a_0 = 1$. One can check that $D_- = H$ is a solution. The second solution can then be obtained from the method of variation of parameters as

$$D_+(a) = \frac{5}{2} \frac{H(a)}{H_0} \Omega_{m0} \int_0^a \frac{da'}{[a'E(a')]^3}, \quad (5.24)$$

where $E(a) = H(a)/H_0$ is given in (3.39). The normalization constant has been chosen so that in a flat matter-dominated Universe, we recover $D_+ = a$.

As an example, in a Universe with $\Lambda = 0$ and $\Omega_m < 1$, it can be shown that the decaying and growing modes are given by

$$D_- = \sqrt{\frac{1+x}{x^3}}, \quad D_+ = 1 + \frac{3}{x} + 3D_-(x) \ln [\sqrt{1+x} - \sqrt{x}], \quad (5.25)$$

where $x \equiv \Omega_m^{-1} - 1$. Note that when $\Omega_m \rightarrow 0$, that is when $x \gg 1$, $D_+ \rightarrow 1$ and $D_- \sim 1/x$ so that the perturbations cease to grow.

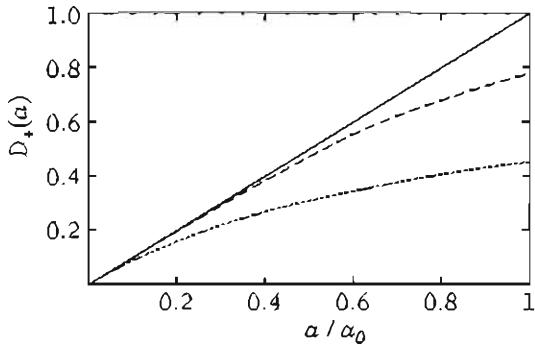


Fig. 5.1 Evolution of the growing mode of the density contrast, $D_+(a)$, as a function of the scale factor, a , for $\Omega_{m0} = 1$ [solid line], $(\Omega_{m0}, \Omega_{\Lambda0}) = (0.3, 0.7)$ [dashed line] and $(\Omega_{m0}, \Omega_{\Lambda0}) = (0.5, 0)$ [dotted line].

In more general cases, (5.24) for D_+ has to be integrated numerically, as illustrated in Fig. 5.1. It has been shown [13] that for a Λ CDM model it is well approximated by

$$D_+ \simeq \frac{5}{2} \frac{a\Omega_m}{\Omega_m^{4/7} - \Omega_\Lambda + (1 + \frac{1}{2}\Omega_m)(1 + \frac{1}{70}\Omega_\Lambda)}. \quad (5.26)$$

Actually, when $K = 0$, it can be checked [14], using the solution (3.46) of the Friedmann equations, that

$$D_+ = {}_2F_1 \left[1, \frac{1}{3}; \frac{11}{6}; -\sinh^2 \left(\frac{3\alpha t}{2} \right) \right] \sinh^{2/3} \left(\frac{3\alpha t}{2} \right),$$

where $\alpha = H_0 \sqrt{\Omega_{\Lambda0}} = \sqrt{\Lambda/3}$ and ${}_2F_1$ is a hypergeometric function.

5.1.2.5 The Hubble parameter from the growth function

Equation (5.23) can also be seen as a function for $E(a)$ once rewritten as

$$\frac{dE^2}{da} + 2 \left[\frac{3}{a} + \frac{d^2 D}{da^2} \left(\frac{dD}{da} \right)^{-1} \right] E^2 - 3 \frac{\Omega_{m0}}{a^5} D \left(\frac{dD}{da} \right)^{-1} = 0.$$

This first-order equation for E^2 is easily integrated to give

$$E^2(z) = -3\Omega_{m0} (1+z)^2 \left(\frac{dD}{dz} \right)^{-2} \int_z^\infty \frac{D}{1+z} \frac{dD}{dz} dz, \quad (5.27)$$

in terms of the redshift $z = a_0/a$. It follows that for the class of Λ CDM models, there exists a rigidity between the background dynamics and the growth of large-scale structure. Such a feature can be useful to test this model (see Chapter 12).

5.1.2.6 Density–velocity relation

Since $\delta \propto D(t)$, (5.15), implies that

$$\theta(\mathbf{x}, t) \equiv \frac{1}{aH} \nabla \cdot \mathbf{u}(\mathbf{x}, t) = -f(\Omega_{m0}, \Omega_{\Lambda0})\delta, \quad (5.28)$$

with

$$f(\Omega_{m0}, \Omega_{\Lambda0}) \equiv \frac{d \ln D(a)}{d \ln a}. \quad (5.29)$$

There exist many analytical approximations for this function. For instance, Ref. [13] suggests

$$f_+(\Omega_{m0}, \Omega_{\Lambda0}) \simeq \Omega_{m0}^{0.6} + \frac{\Omega_{\Lambda0}}{70} \left(1 + \frac{1}{2}\Omega_{m0}\right).$$

Note that θ represents the local fluctuations of the Hubble constant. This is an observable quantity, independent of the value of the Hubble constant. Using (5.13), it follows that the velocity field has the general form

$$\mathbf{u} = -\frac{2}{3} \frac{f_+(\Omega_{m0}, \Omega_{\Lambda0})}{aH\Omega_m} \nabla\Phi + \nabla \times \mathbf{U}, \quad (5.30)$$

where \mathbf{U} is an arbitrary vector function.

Considering again the example of a flat Universe with no cosmological constant, the solution (5.22) implies that

$$f_+ = 1, \quad \mathbf{u}_+ = -t^{1/3} \nabla\Phi, \quad f_- = -\frac{3}{2}, \quad \mathbf{u}_- = t^{-1/3} \nabla\Phi. \quad (5.31)$$

Thus, the growing mode corresponds to a velocity field falling towards the potential well, which does indeed increase δ . The decaying mode corresponds to a configuration where the fluid escapes from the potential wells and tends to erase the initial fluctuations.

Since the velocity field can now be observed [15], it opens the possibility of testing the gravitational dynamics and provides a way to measure the cosmological parameters.

5.1.2.7 Gravitational potential

In the linear regime, and in the Newtonian approach we have followed in this section, the gravitational potential can be obtained from the Poisson equation (5.5) so that it behaves as $\Phi \propto D_+(t)/a(t)$. It is thus constant in an Einstein–de Sitter space-time ($\Omega_m = 1$, $K = 0$ and $\Lambda = 0$); see (5.22).

The typical amplitude of the density fluctuation is usually characterized by σ_8 ; see definition (5.45) below. The Poisson equation (5.5) allows us to estimate the typical amplitude of the gravitational potential as

$$\sigma_\Phi = \frac{3}{2} \Omega_{m0} \left(\frac{8 \text{ Mpc}}{H_0^{-1}} \right)^2 \sigma_8.$$

It follows that $\sigma_\Phi \sim 10^{-5}$ if $\sigma_8 \sim 1$. This estimate has two consequences. First, it implies that Φ can be treated to first order in the perturbations even if the density

contrast has entered the non-linear regime. Second, if the power spectrum is (almost) scale invariant, as predicted by inflation (see Chapter 8), then this amplitude of the gravitational potential is its typical amplitude on all scales.

5.1.3 Predictions and observables

We have characterized the density and velocity fields of the cosmic matter fluid. The density field results from the gravitational collapse of primordial density perturbations. In this book, we will try to present the models capable of generating these primordial fluctuations. It is, however, important to discuss some aspects of the comparison of the predictions we will make with the observed Universe.

5.1.3.1 The need for a statistical approach

Note first of all that the primordial fluctuations are not directly observable. Only different objects, observed at different epochs of their evolution, are accessible. For instance, although $\epsilon_{\pm}(x)$, the field of primordial fluctuations, is not directly accessible, we can observe the large structures of the Universe (galaxies, clusters...), and yet we can only measure a limited set of properties of these objects spatial distribution.

As will be seen later, models of the primordial Universe allow us to predict the statistical properties of the primordial fluctuations of which we will be able to deduce the distribution of some observables such as δ . The density field, as well as all the other fields, are thus now to be considered as stochastic variables. Every statistical prediction will hence depend on the statistical properties of the initial perturbations and on their dynamical evolution. The Universe is thus modelled as a stochastic realization of a statistical set of possibilities.

5.1.3.2 Statistical description

Given the above, we will thus be dealing with classical random fields, such as the density contrast, for instance, and compute some of their statistical properties such as the correlation function

$$\xi(x, r) = \langle \delta(x)\delta(x + r) \rangle. \quad (5.32)$$

Let us draw the implication of the Copernican principle when applied statistically to the fields living in a Friedmann-Lemaître space-time.

Statistical homogeneity means that all the moments of a random field or the joint distribution probabilities $p[\delta(r_1), \delta(r_2), \dots]$ remain unchanged under the action of any space translation. So these probabilities only depend on the relative separations, such as $r_1 - r_2$. Statistical isotropy, on the other hand, demands that these quantities be also invariant under the action of rotations. Therefore, the probabilities eventually depend only on the modulus of the relative separations, namely e.g. $r_{12} = |r_1 - r_2|$. This hypothesis can be considered as the statistical version of the cosmological principle, but it is only an hypothesis and needs to be tested (as should the Copernican principle itself).

Assuming statistical isotropy and homogeneity, the correlation function (5.32) then only depends on $r \equiv |r|$, i.e.

$$\xi(x, r) = \xi(r).$$

It is convenient to expand $\delta(\mathbf{r})$ into Fourier modes as

$$\delta(\mathbf{r}) = \int \frac{d^3 k}{(2\pi)^{3/2}} \delta_{\mathbf{k}} e^{i\mathbf{k} \cdot \mathbf{r}}, \quad (5.33)$$

and since we consider a real-valued field, the coefficients $\delta_{\mathbf{k}}$ satisfy the conjugation relation

$$\delta_{\mathbf{k}}^* = \delta_{-\mathbf{k}}. \quad (5.34)$$

The random field is thus completely characterized by the statistical properties of the random variables $\delta_{\mathbf{k}}$. For instance

$$\langle \delta_{\mathbf{k}} \delta_{\mathbf{k}'} \rangle = \int \frac{d^3 x}{(2\pi)^{3/2}} \frac{d^3 r}{(2\pi)^{3/2}} \langle \delta(x) \delta(x + \mathbf{r}) \rangle e^{-i(\mathbf{k} + \mathbf{k}') \cdot \mathbf{x} - i\mathbf{k}' \cdot \mathbf{r}}, \quad (5.35)$$

which, after integrating over \mathbf{x} , gives

$$\langle \delta_{\mathbf{k}} \delta_{\mathbf{k}'} \rangle = \delta^{(3)}(\mathbf{k} + \mathbf{k}') P_{\delta}(k), \quad (5.36)$$

$\delta^{(3)}$ being the Dirac distribution, not to be confused with the density perturbation. The power spectrum P_{δ} is related to the correlation function by

$$P_{\delta}(k) = \int d^3 r \xi(r) e^{i\mathbf{k} \cdot \mathbf{r}}, \quad (5.37)$$

a relation that can be inverted to yield an expression for the correlation function, $\xi(r)$, in terms of the power spectrum, namely

$$\xi(r) = \int \frac{d^3 k}{(2\pi)^3} P_{\delta}(k) e^{-i\mathbf{k} \cdot \mathbf{r}} = \int_0^\infty \frac{dk}{2\pi^2} k^2 P_{\delta}(k) \frac{\sin kr}{kr}. \quad (5.38)$$

5.1.3.3 Link with observable quantities and smoothing

Quantities such as $\xi(r)$ cannot actually be measured as we only observe a finite part of a single realization of the stochastic process. So, operationally, only the spatial average is accessible

$$\xi_{\text{obs}}(r) = \frac{1}{V} \int_V d^3 r' \delta_{\text{obs}}(\mathbf{r}') \delta_{\text{obs}}(\mathbf{r} + \mathbf{r}'), \quad (5.39)$$

where δ_{obs} is the observed field density. For any statistical observable, let us say O , one must construct an estimator, \hat{O} . The probability of measuring the value \hat{O} in a cosmological survey (e.g. galaxy catalogue) given a theory is called the *cosmic distribution function*, $\mathcal{P}(\hat{O})$, and the ensemble average of \hat{O} is simply

$$\langle \hat{O} \rangle = \int \hat{O} \mathcal{P}(\hat{O}) d\hat{O}.$$

Most estimators are non-linear and are thus biased, in the sense that their ensemble average differs from the real value O . The cosmic bias b_O , defined as

$$b_O \equiv \frac{\langle \hat{O} \rangle - O}{O},$$

does not vanish in general, except when the size of the survey tends to infinity, since in that case one measures the actual value of the observable. A good estimator must have a minimum cosmic bias and also minimize the cosmic error, which is obtained by computing the variance of \mathcal{P} .

This description is still idealized and simplistic. Indeed, what does it mean to measure $\delta_{\text{obs}}(r)$? For instance, a catalogue of galaxies can give us access to the measurement of the mean number of galaxies per volume V so that an estimate of a continuous field should be inferred from the observation of the number of objects in a given volume. This gives an estimate of the average density in this volume, namely

$$\delta_{\text{obs},R}(x) = \int_V d^3y \delta_{\text{obs}}(y) W_R(x-y), \quad (5.40)$$

thereby defining the (observational) window function W_R . This function must satisfy a normalization condition

$$\int W_R(x) d^3x = 1, \quad (5.41)$$

as well as the condition

$$\int W_R(x) x^2 d^3x = R^2, \quad (5.42)$$

which guarantees that its mean width is finite and of characteristic size R . The archetypes of such ‘window’ functions are the ‘top-hat’ and Gaussian filters. Note that if the Fourier components of this filter are denoted by $\widehat{W}_R(k)$, then the Fourier coefficients of the filtered field are given by

$$\delta_R(k) = \delta_k \widehat{W}_R(k), \quad (5.43)$$

since (5.40) is a convolution.

The cosmic error and the cosmic bias have three main contributions (for a discussion of these effects, see Refs. [12, 16]):

- (1) *Finite-volume effects*, due to the fact that there are approximatively $N \sim V/L^3$ independent volumes of size L in a survey of volume V . In particular, the mean value of the density is not always well determined. These effects are proportional to the average of the 2-point correlation function over the survey and are usually referred to under the name of *cosmic variance*.
- (2) *Edge effects*, related to the geometry of the catalogue and to the fact that the estimators in general give less weight to the data on the edge of the survey.
- (3) *Discreteness effects*, related to the fact that the theoretical field is assumed continuous while the observed one (e.g. galaxies) is discrete. These effects scale as the inverse of the number of objects in the survey to some power.

Thus, for any physical quantity $Q(r)$ we would like to predict the probability distribution function (PDF) of the filtered quantity $Q_R(r)$ (we restrict ourselves to the one-point function for the sake of simplicity). This PDF, $p(Q, R)dQ$, represents the

probability to lie between Q and $Q + dQ$. For example, if Q is the number of galaxies and the filter a top-hat, then the PDF of Q_R will give the probability of finding N galaxies within a volume of radius R . Models such as inflation predict that these initial fluctuations are Gaussian, i.e.

$$p(Q_i, R) = \frac{1}{\sqrt{2\pi}\sigma_i(R)} \exp\left[-\frac{Q_{iR}^2}{2\sigma_i^2(R)}\right], \quad (5.44)$$

where $\sigma_i^2(R) = \langle Q_{iR}^2 \rangle$ is the initial variance. The second moment completely characterizes this distribution, which we would like to compare with that of Q_R today. There are thus two effects to consider:

- the evolution of the modes can modify the initial spectrum by a transfer function that may depend on k ,
- non-linear evolution can modify the initial statistics.

There is a third effect that will not be discussed in this book, which comes from the fact that only luminous objects are observed. It should therefore be checked that luminous matter is a good tracker of the total matter density. This is at the origin of the notion of *biased measure* [12, 17–19].

The rest of this chapter is dedicated to the first aspect that requires a detailed analysis of the evolution of perturbations in the relativistic framework. We briefly illustrate the second aspect in the next section.

5.1.4 Towards the non-linear regime

The Newtonian description allows us to address the behaviour of perturbations in the non-linear regime. This aspect cannot be developed further in this book, but we refer the reader to the review [12] for a very detailed study.

Whatever its initial value, δ grows and, given enough time, eventually becomes of order unity. The time at which this happens depends on the initial amplitude of the fluctuation and on the growth factor, that is on the cosmological parameters. As we will see, the normalization of the initial power spectrum is such that the (top-hat) variance of the density field in a sphere of radius $R_8 = 8h^{-1}$ Mpc is of order unity

$$\sigma_8^2 \equiv \left\langle \left(\frac{3}{4\pi R_8^3} \int_{|x| \leq R_8} \delta(x) d^3x \right)^2 \right\rangle \sim 1. \quad (5.45)$$

The length scale R_8 therefore characterizes the scale below which non-linearities cannot be neglected, i.e. the scale below which the density contrast is too large for the linear approximation to be applicable.

To take the non-linear evolution into account, the density field can be expanded perturbatively as

$$\delta = \delta^{(1)} + \delta^{(2)} + \dots,$$

where the term $\delta^{(n)}$ is of the order of ε^n in the initial density field [$\varepsilon(x) \ll 1$]. The first order corresponds to the solutions previously found

$$\delta^{(1)} = D(t)\varepsilon(x), \quad (5.46)$$

whereas, to second order, the evolution equation (5.16) reduces to

$$\ddot{\delta}^{(2)} + 2H\dot{\delta}^{(2)} = 4\pi G_N \rho \left[\delta^{(1)} \right]^2 + \frac{1}{a^2} \nabla \delta^{(1)} \cdot \nabla \Phi + \frac{1}{a^2} \partial_i \partial_j \left[u_{(1)}^i u_{(1)}^j \right]. \quad (5.47)$$

Using the 1st-order solutions, one can solve this equation to obtain [20]

$$\delta^{(2)} = D^2(t) \left\{ \frac{2}{3} [1 + \kappa(t)] \varepsilon^2(x) + \nabla \varepsilon \cdot \nabla \Phi + \left[\frac{1}{2} - \kappa(t) \right] \sigma^2 \right\}, \quad (5.48)$$

with

$$\sigma^2 = \sigma_{ij} \sigma^{ij}, \quad \sigma_{ij} = \partial_i \partial_j \Phi - \frac{1}{3} \delta_{ij} \Delta \Phi.$$

The function $\kappa(t)$ is slowly varying, and to a good approximation is given by [21]

$$\kappa \approx \frac{3}{14} \Omega_m^{-2/63},$$

when the cosmological constant vanishes.

5.1.4.1 Application: Skewness

In order to illustrate the effect of non-linear dynamics, let us consider the third-order moment of the density field

$$\langle \delta^3 \rangle = \left\langle (\delta^{(1)})^3 \right\rangle + \left\langle (\delta^{(1)})^2 \delta^{(2)} \right\rangle + \mathcal{O}(\varepsilon^5),$$

where the first term in this expansion vanishes for Gaussian initial conditions. We see that the non-linear evolution induces a departure from Gaussianity of the density field.

As discussed previously, one should consider the smoothed density field, and at the lowest order, we obtain

$$\langle \delta_R^3 \rangle = D^2(t) \int \frac{d^3 k}{(2\pi)^3} \frac{d^3 k'}{(2\pi)^3} P_\delta(k) P_\delta(k') \widehat{W}_R(k) \widehat{W}_R(k') \widehat{W}_R(|k+k'|) J(k, k'), \quad (5.49)$$

with

$$J(k, k') = \frac{1}{2} + \kappa + \frac{k'}{k} \left(\frac{k \cdot k'}{kk'} \right) + \left(\frac{1}{2} - \kappa \right) \left(\frac{k \cdot k'}{kk'} \right)^2.$$

Neglecting the filter and assuming that the spectrum behaves as a power law $P(k) \propto k^n$, we obtain [21]

$$S_3(0) \equiv \frac{\langle \delta^3 \rangle}{\langle \delta^2 \rangle^2} = 4 + 4\kappa - (n+3) \simeq \frac{34}{17} \left(\Omega_m^{-2/63} - 1 \right) - (n+3). \quad (5.50)$$

For a top-hat filter, this result is modified [22] to

$$S_3(R) = S_3(0) + \frac{d \ln \sigma_\delta^2(R)}{d \ln R}, \quad (5.51)$$

where $\sigma_\delta(R)$ is the variance of the density field, filtered at the scale R . Such a relation can be compared with observations (see, for instance, Ref. [23]), e.g. to show that large-scale structures are compatible with Gaussian initial conditions.

5.2 Gauge invariant cosmological perturbation theory

The Newtonian approach is valid for sub-Hubble perturbations and is relevant for the description of the evolution of perturbations in the late Universe. However, the gravitational dynamics must be described in a relativistic framework, particularly for long-wavelength modes. The following section is one of the cornerstones of contemporary theoretical cosmology. In particular, we will present a derivation of the equations generalizing (5.1), (5.2) and (5.5).

There exist two formalisms for cosmological perturbation theory. We choose to present an approach in which we introduce arbitrary parameterizations of the perturbations from which we will then construct gauge invariant variables. This method, introduced by Bardeen [24, 25], is the most common one and various aspects are detailed in many reviews [26–29]. For the approach that we do not develop, Refs. [30–32] can be consulted. Both approaches lead to equivalent physical results.

5.2.1 Perturbed space-time

5.2.1.1 Metric of the perturbed space-time

At linear order, we decompose the space-time metric according to

$$ds^2 = a^2(\eta) \left[-(1 + 2A)d\eta^2 + 2B_i dx^i d\eta + (\gamma_{ij} + h_{ij})dx^i dx^j \right], \quad (5.52)$$

where the small quantities A , B_i and h_{ij} are unknown functions of space and time to be determined from the Einstein equations. Writing this metric as

$$g_{\mu\nu} = \bar{g}_{\mu\nu} + \delta g_{\mu\nu}, \quad (5.53)$$

where $\bar{g}_{\mu\nu}$ is the Friedmann–Lemaître metric, the inverse metric is then given by

$$g^{\mu\nu} = \bar{g}^{\mu\nu} + \delta g^{\mu\nu}, \quad \delta g^{\mu\nu} = -\bar{g}^{\mu\sigma} \bar{g}^{\nu\lambda} \delta g_{\sigma\lambda}, \quad (5.54)$$

to first order in the perturbations. In what follows, h will denote the trace of the spatial perturbations

$$h \equiv h_{ij} \gamma^{ij},$$

and we recall that ∇_μ is the covariant derivative associated with $g_{\mu\nu}$. The Christoffel symbols, Riemann, Ricci and Einstein tensors and the scalar curvature associated with this metric are given explicitly in Appendix C.

5.2.1.2 SVT decomposition

In what follows, we will perform a so-called scalar–vector–tensor decomposition [24] (denoted by SVT in the following) of the perturbation variables. This is, in a cosmological background, the same decomposition as that introduced in Chapter 1.

This decomposition is a generalization of the well-known fact that any vector field can be decomposed as the sum of the gradient of a scalar and a divergenceless vector as

$$B^i = D^i B + \bar{B}^i \quad \text{with} \quad D^i \bar{B}_i = 0. \quad (5.55)$$

For a velocity field, B will be the potential and \bar{B}_i the vorticity. One sees that the 3 components of the vector have been split into 1 scalar (B) and 2 vector (\bar{B}_i) components.

In an analogous way, any rank 2 symmetric tensor can be decomposed as

$$h_{ij} = 2C\gamma_{ij} + 2D_i D_j E + 2D_{(i}\bar{E}_{j)} + 2\bar{E}_{ij} \quad \text{with} \quad D_i \bar{E}^{ij} = 0, \quad \bar{E}_i^i = 0. \quad (5.56)$$

Equation (5.54) implies that the indices of these quantities are raised and lowered with respect to the unperturbed spatial metric, γ_{ij} , for example, $\bar{B}^i = \gamma^{ij} \bar{B}_j$. We adopt the convention in which any ‘barred’ quantity is divergenceless (and traceless if it has two indices). The 6 components of h_{ij} have thus been split into 2 scalar (C and E), 2 vector (\bar{E}_j) and 2 tensor \bar{E}_{ij} components.

Note that the quantities B and E are not unique. In fact, B is a solution of $\Delta B = D_i B^i$ (the Laplacian being defined by $\Delta \equiv D_i D^i$), whose solution is unique only if some boundary conditions are imposed.

The 10 degrees of freedom of the metric have thus been decomposed as

- 4 scalars: A , B , C and E , corresponding to 4 degrees of freedom,
- 2 vectors: \bar{B}^i and \bar{E}^i corresponding to $2 \times (3 - 1) = 4$ degrees of freedom,
- 1 tensor: \bar{E}^{ij} , corresponding to $3 \times 2 - 1 - 3 = 2$ degrees of freedom.

The advantages of this decomposition lie in the fact that, to leading order, the three types of perturbations are decoupled and can thus be studied separately.

5.2.1.3 The gauge problem

To introduce the perturbed space-time metric, we have made the hypothesis (5.53) that the metric $g_{\mu\nu}$ was ‘close’ to that of the Friedmann–Lemaître $\bar{g}_{\mu\nu}$.

In field theory, space-time is usually fixed and once the system of coordinates is chosen, one can define the perturbation of any quantity $Q(\mathbf{x}, t)$ as

$$\delta Q(\mathbf{x}, t) = Q(\mathbf{x}, t) - \bar{Q}(\mathbf{x}, t),$$

where $\bar{Q}(\mathbf{x}, t)$ represents the unperturbed configuration. At any point of space-time, the quantity $Q(\mathbf{x}, t)$ can be compared with $\bar{Q}(\mathbf{x}, t)$ at the same point.

In general relativity, there exist an important difference with this standard approach. Indeed, we would like to compare two different space-times and make the hypothesis that $(\mathcal{M}, g_{\mu\nu})$ is ‘close’ to a Friedmann–Lemaître Universe $(\bar{\mathcal{M}}, \bar{g}_{\mu\nu})$. In making this comparison, there is some arbitrary freedom related to the way of identifying the points of these two space-times. Indeed, an isomorphism ψ should be introduced to relate the points of the two space-times (see Fig. 5.2). There is actually no ‘natural’ identification between the two spaces, and there is therefore a freedom of choice in the identification of the points in each spaces. This implies that there exist some unphysical degrees of freedom related to the choice of the coordinate systems on the two manifolds.

The decomposition (5.52) implicitly assumes that a system of coordinates has been chosen in the perturbed space. Any change of coordinates will modify the form of the metric coefficients. To illustrate the effect of such a change of coordinates, let us consider an unperturbed Friedmann–Lemaître space-time, as in Fig. 5.2, and let

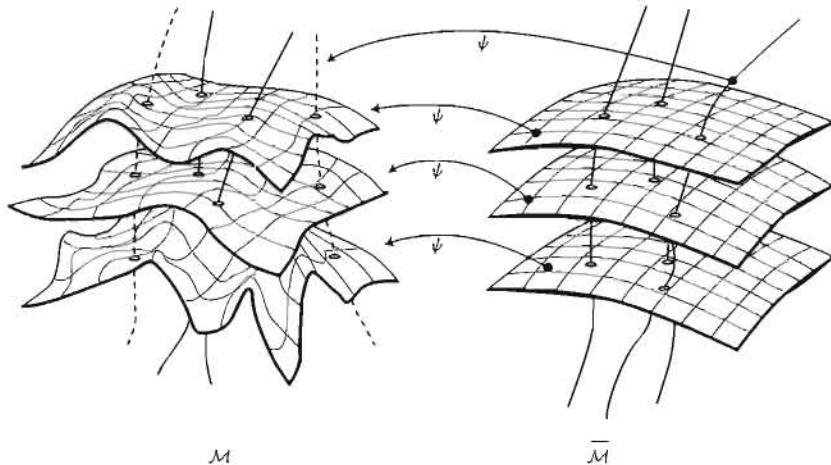


Fig. 5.2 Any perturbed quantity is defined via a mapping between the Friedmann–Lemaître space-time, $\bar{\mathcal{M}}$, and the perturbed space-time \mathcal{M} .

us perform an arbitrary space and time-dependent change of coordinates $x^i \rightarrow y^i = x^i - \xi^i(x^j, \eta)$. We then get

$$ds^2 = a^2(\eta) [-d\eta^2 + 2\xi'_i dy^i d\eta + (\gamma_{ij} + 2D_{(i}\xi_{j)}) dy^i dy^j].$$

There appear two artificial terms $B_i = \xi'_i$ and $E_i = \xi_i$ that do not correspond to metric fluctuations, since this space is still a Friedman–Lemaître space, but to a choice of the system of coordinates. The dependence of the quantities that are intrinsic to the manifold in which the perturbations evolve should thus be separated from the artificial ones, related to the arbitrary choice of a particular coordinate system.

5.2.1.4 Gauge invariant variables

In order to extract the physical degrees of freedom, let us consider an active transformation of the coordinate system defined by a vector field ξ . The coordinates of any point change according to

$$x^\mu \rightarrow x^\mu - \xi^\mu, \quad (5.57)$$

where the displacement ξ^μ is decomposed into 2 scalar degrees of freedom (T and L) and 2 vector degrees of freedom (\bar{L}^i , which is divergenceless $D_i \bar{L}^i = 0$) as

$$\xi^0 = T, \quad \xi^i = L^i = D^i L + \bar{L}^i. \quad (5.58)$$

Under this change of coordinates, the metric transforms as [see (1.57)]

$$g_{\mu\nu} \rightarrow g_{\mu\nu} + \mathcal{L}_\xi g_{\mu\nu}.$$

To first order in the perturbations, this is

$$\delta g_{\mu\nu} \rightarrow \delta g_{\mu\nu} + \mathcal{L}_\xi \bar{g}_{\mu\nu} = \delta g_{\mu\nu} + 2\nabla_{(\mu} \xi_{\nu)}, \quad (5.59)$$

which in turn implies that the metric perturbation variables (5.52) transform as¹

$$A \rightarrow A + T' + \mathcal{H}T \quad (5.60)$$

$$B_i \rightarrow B_i - D_i T + L'_i$$

$$h_{ij} \rightarrow h_{ij} + D_i L_j + D_j L_i + 2\mathcal{H}T \gamma_{ij}. \quad (5.61)$$

Using the scalar-vector-tensor decomposition (5.55) and (5.56) we get

| | | |
|--|---|--|
| $A \rightarrow A + T' + \mathcal{H}T,$ $B \rightarrow B - T + L',$ $C \rightarrow C + \mathcal{H}T,$ $E \rightarrow E + L,$ | $\bar{B}^i \rightarrow \bar{B}^i + \bar{L}^{i\nu},$ $\bar{E}^i \rightarrow \bar{E}^i + \bar{L}^i,$ | $\bar{E}_{ij} \rightarrow \bar{E}_{ij},$ |
|--|---|--|

(5.62)

respectively, for the S, V and T modes. These transformations are similar to the gauge transformations encountered, for instance, in electromagnetism. Therefore, one may reduce the arbitrariness by imposing some additional conditions on the variables: in other words, and again as in electromagnetism, one needs to fix the gauge before doing any actually relevant calculation.

As stressed in the previous section, it is interesting to define gauge-independent quantities. To do so, notice that some combinations of the previous quantities do not depend on L^i and T , for instance,

$$\Psi \equiv -C - \mathcal{H}(B - E'), \quad (5.63)$$

$$\Phi \equiv A + \mathcal{H}(B - E') + (B - E')', \quad (5.64)$$

$$\bar{\Phi}^i \equiv \bar{E}^{i\nu} - \bar{B}^i, \quad (5.65)$$

$$\bar{E}^{ij}. \quad (5.66)$$

We initially had 10 perturbation variables ($A, B, C, E, \bar{E}^i, \bar{B}^i, \bar{E}^{ij}$) and 4 gauge degrees of freedom (T, L, \bar{L}^i) that can be absorbed into the definition of the gauge invariant variables, then described by $10 - 4 = 6$ degrees of freedom parameterized, for instance, by $\Psi, \Phi, \bar{\Phi}^i$ and \bar{E}^{ij} . These quantities can be considered as the 'real' space-time perturbations in the sense that they cannot be removed by any change of coordinates.

¹For completeness, we give the expressions of the Lie derivatives. They are easily obtained from the expressions of the Christoffel symbols of the Friedmann-Lemaître space-time and from $\xi_\mu = a^2(-T, L_i)$. They are given by

$$\mathcal{L}_\xi \hat{g}_{00} = -2a^2(T' + \mathcal{H}T),$$

$$\mathcal{L}_\xi \hat{g}_{0i} = a^2(-\partial_i T + L'_i).$$

$$\mathcal{L}_\xi \hat{g}_{ij} = a^2(D_i L_j + D_j L_i + 2\mathcal{H}T \gamma_{ij}).$$

Other gauge invariant quantities could have been constructed, such as, for instance,

$$X = A - C - \left(\frac{C}{\mathcal{H}} \right)', \quad (5.67)$$

but they can always be expressed in terms of the two functions, Φ and Ψ . For instance, in the case of X defined just above,

$$X = \Psi + \Phi + \left(\frac{\Psi}{\mathcal{H}} \right)', \quad (5.68)$$

as can straightforwardly be checked.

5.2.1.5 Stewart–Walker lemma

Just as the metric transforms as (5.59) under a change of coordinates of the form (5.57), any scalar quantity δQ transforms as

$$\delta Q \rightarrow \delta Q + \mathcal{L}_\xi \bar{Q},$$

to first order in perturbations. Since each vector field ξ generates a gauge transformation, we can conclude that the only gauge invariant quantities are the ones for which

$$\mathcal{L}_\xi \bar{Q} = 0 \quad \forall \xi.$$

This result is known under the name of the Stewart–Walker lemma [33].

Note an important corollary of this result. Since all relativistic equations are covariant by definition, they can always be written in the form $Q = 0$, where Q is a tensor field. It is thus always possible, to first order in perturbations, to write all the relevant equations in terms of gauge invariant quantities only. Such a treatment would ensure that no gauge, i.e. unphysical, degree of freedom, is misleadingly used.

5.2.2 Description of matter

In this section, we introduce various quantities that come into the description of a perfect fluid for which the energy-momentum tensor in an unperturbed space is of the form

$$T_{\mu\nu} = (\rho + P)u_\mu u_\nu + Pg_{\mu\nu},$$

as was seen in Chapter 3.

5.2.2.1 The energy-momentum tensor of a perturbed fluid

The energy-momentum tensor of this perturbed fluid takes the general form

$$\delta T_{\mu\nu} = (\delta\rho + \delta P)\bar{u}_\mu \bar{u}_\nu + \delta P \bar{g}_{\mu\nu} + 2(\rho + P)\bar{u}_{(\mu} \delta u_{\nu)} + P\delta g_{\mu\nu} + a^2 P \pi_{\mu\nu}, \quad (5.69)$$

where $u^\mu = \bar{u}^\mu + \delta u^\mu$ is the four-velocity of a comoving observer, satisfying $u_\mu u^\mu = -1$. The normalization condition of \bar{u}^μ (to zeroth order) provides the solution

$$\bar{u}^\mu = a^{-1} \delta_0^\mu, \quad \bar{u}_\mu = -a \delta_\mu^0,$$

and since the norm of $\bar{u}^\mu + \delta u^\mu$ should also be equal to -1 , we infer that $2\bar{u}^\mu \delta u_\mu + \delta g_{\mu\nu} \bar{u}^\mu \bar{u}^\nu = 0$, and thus that $\delta u_0 = -Aa$. We then write $\delta u^i \equiv v^i/a$, so that

$$\delta u^\mu = a^{-1}(-A, v^i), \quad \delta u_\mu = a(-A, v_k + B_k), \quad (5.70)$$

and we decompose v_i into a scalar and a tensor part according to

$$v_i = D_i v + \bar{v}_i. \quad (5.71)$$

In the decomposition (5.69), $\pi_{\mu\nu}$ is the anisotropic stress tensor and characterizes the difference between the perturbed fluid and a perfect fluid. It can be chosen to be traceless ($g^{\mu\nu} \pi_{\mu\nu} = 0$), since its trace can be absorbed into a redefinition of the isotropic pressure, δP . It is also symmetric, since the energy-momentum tensor is, and can be chosen orthogonal to u^μ , i.e. $u^\mu \pi_{\mu\nu} = 0$. This implies that, without lack of generality, one can set $\pi_{00} = \pi_{0i} = 0$.

The anisotropic stress tensor then consists only of a spatial part, which can be decomposed into scalar, vector and tensor parts as

$$\pi_{ij} = \Delta_{ij} \bar{\pi} + D_{(i} \bar{\pi}_{j)} + \bar{\pi}_{ij}, \quad (5.72)$$

where the operator Δ_{ij} is defined as

$$\Delta_{ij} \equiv D_i D_j - \frac{1}{3} \gamma_{ij} \Delta. \quad (5.73)$$

We infer from (5.70), (5.71) and (5.72) that the components of the perturbation of the energy-momentum tensor (5.69) are

$$\boxed{\delta T_{00} = \rho a^2 (\delta + 2A),} \quad (5.74)$$

$$\boxed{\delta T_{0i} = -\rho a^2 [(1+w)(D_i v + \bar{v}_i) + D_i B + \bar{B}_i],} \quad (5.75)$$

$$\boxed{\delta T_{ij} = P a^2 \left(h_{ij} + \frac{\delta P}{P} \gamma_{ij} + \pi_{ij} \right),} \quad (5.76)$$

where we recall that $\delta \equiv \delta\rho/\bar{\rho}$ is the density contrast.

It will be useful to introduce the entropy perturbation, Γ , defined by the relation

$$\delta P = c_s^2 \delta \rho + P \Gamma, \quad (5.77)$$

which takes the form

$$w\Gamma = \frac{1}{\rho} (\delta P - c_s^2 \delta \rho). \quad (5.78)$$

For an adiabatic perturbation, $\delta P/\delta \rho = c_s^2$ and hence $\Gamma = 0$ and we introduce the notation used in the literature

$$\delta P_{\text{nad}} = P \Gamma, \quad (5.79)$$

where the index ‘_{nad}’ stands for non-adiabatic.

5.2.2.2 Gauge invariant quantities

Upon performing a change of coordinates of the form (5.57), scalar and vectors transform as

$$\delta Q \rightarrow \delta Q + \mathcal{L}_\xi \bar{Q}, \quad \mathcal{L}_\xi \bar{Q} = \xi^\alpha \partial_\alpha \bar{Q} = T \bar{Q}',$$

and

$$\delta u^\mu \rightarrow \delta u^\mu + \mathcal{L}_\xi \bar{u}^\mu, \quad \mathcal{L}_\xi \bar{u}^\mu = \xi^\alpha \partial_\alpha \bar{u}^\mu - u^\alpha \partial_\alpha \xi^\mu,$$

so that it follows that, after SVT decomposition, δP , $\delta \rho$, v and \bar{v}_i transform as

$$\begin{aligned} \delta \rho &\rightarrow \delta \rho + \rho' T, & v &\rightarrow v - L', \\ \delta P &\rightarrow \delta P + P' T, & \bar{v}_i &\rightarrow \bar{v}_i - \bar{L}'_i. \end{aligned}$$

(5.80)

The quantities π , $\bar{\pi}_i$ and $\bar{\pi}_{ij}$ are gauge invariant from the Stewart–Walker lemma, since the unperturbed part of the anisotropic stress tensor must vanish to be compatible with the perfect fluid hypothesis imposed by the symmetries of the Friedmann–Lemaître space-time. One can also check that, as expected again from the Stewart–Walker lemma, Γ is gauge invariant since $\Gamma \rightarrow \Gamma + (P' - c_s^2 \rho')T = \Gamma$.

As in the case of the metric perturbations, one can likewise define the gauge invariant quantities associated with the quantities (5.80). Different choices are possible, and we define

$$\delta^N = \delta + \frac{\rho'}{\rho}(B - E'), \quad (5.81)$$

$$\delta^F = \delta - \frac{\rho'}{\rho} \frac{C}{H}, \quad (5.82)$$

$$\delta^C = \delta + \frac{\rho'}{\rho}(v + B), \quad (5.83)$$

$$V = v + E', \quad (5.84)$$

$$\bar{V}_i = \bar{v}_i + \bar{B}_i, \quad (5.85)$$

$$\bar{W}_i = \bar{v}_i + \bar{E}'_i. \quad (5.86)$$

The pressure perturbations δP^N , δP^F and δP^C are defined in an identical way. The different gauge invariant variables are related to each other and one can check that

$$\delta^F = \delta^N + \frac{\rho'}{\rho} \frac{\Psi}{H}, \quad \delta^C = \delta^N + \frac{\rho'}{\rho} V, \quad \bar{W}_i = \bar{V}_i + \bar{\Phi}_i. \quad (5.87)$$

In all these relations, we recall that $\delta = \delta\rho/\rho$ is the density contrast.

5.2.3 Choosing a gauge

From the Stewart–Walker lemma, we know that any perturbation equation can be written in terms of gauge invariant variables only. Indeed, such an equation always takes the form $Q = 0$ and is, by construction, linear in the first-order perturbation variables. One method for obtaining these gauge invariant equations is thus to write

the equations in an arbitrary gauge and to regroup the terms to make the gauge invariant variables appear.

A more subtle way is to work in a given gauge from the beginning so that the variables in that gauge reduce to gauge invariant variables. For instance, the longitudinal gauge is fixed by the condition $E = B^i = 0$, so that $C = -\Psi$, $A = \Phi$ The equations will then be derived directly in terms of gauge invariant variables. We can then change gauge or re-express the equations in terms of the initial gauge dependent perturbations by using, e.g., (5.63)–(5.66).

Let us stress an important, and often confusing, point. Although the calculations can be made in any gauge, one in general needs to perform a gauge transformation to relate the results of these calculations with observable quantities. In fact, the gauge invariant quantities do not have a priori any direct physical interpretation and one needs to find a gauge where the perturbation variables reduce to the gauge invariant quantities. For instance, δ^N can only be interpreted as the density contrast if it is measured by an observer using this gauge.

Note also that locally, all systems of coordinates are equivalent. Therefore, we only expect to find differences between the different gauges on large scales, or, more precisely, above some yet-unknown characteristic scale. We will see shortly that this is indeed the case and that the scale in question is fixed by the Hubble radius.

A final cautionary point at this level: if, for some reason, a supposedly small (perturbative) quantity, calculated in a fixed gauge, becomes large, while all quantities calculated in a different gauge remain small, this does not necessarily mean that the first-order perturbation theory breaks down, but neither does it mean the opposite! In fact, it merely implies that the coordinate change to get to this gauge is no longer infinitesimal. One then needs to go on to second-order calculations, at least: if this second-order correction happens to be smaller than the first order in all gauges, then it is likely that the particular gauge used to begin with was pathological at this point. This, however, really needs be checked before any physical conclusion be reached.

We now turn to the extremely useful task of presenting the various gauges frequently used, before deriving the equations of evolution for the cosmological perturbations.

5.2.3.1 Newtonian or longitudinal gauge

This gauge is the one where the scalar part of the perturbed metric is diagonal, namely

$$B = 0 \quad \text{and} \quad E = 0. \quad (5.88)$$

Given these conditions, the vector part cannot be diagonal, and one can arbitrarily choose to impose the condition

$$\bar{B}_i = 0, \quad (5.89)$$

which fixes the gauge completely since one can pass from an arbitrary gauge to this one with the transformation (5.58) with parameters defined by

$$T = B - E', \quad L = -E \quad \text{and} \quad \bar{L}'_i = -\bar{B}_i. \quad (5.90)$$

In this gauge, the expansion is 'seen' as isotropic. The quantity Ψ represents the curvature perturbation of constant time hypersurfaces [see (C.31)] and it is this potential

that arises in the Poisson equation, i.e. the one that reduces to the Newtonian potential on small scales, hence the gauge name.

In this gauge, one finds the remaining perturbed quantities

$$A = \Phi, \quad C = -\Psi, \quad \delta = \delta^N, \quad \delta P = \delta P^N, \quad v = V \quad \text{and} \quad \bar{v}_i = \bar{V}_i, \quad (5.91)$$

in terms of the gauge invariant variables.

5.2.3.2 Flat-slicing gauge

The flat-slicing gauge is defined by imposing that the scalar part of the curvature perturbation of the spatial sections vanishes. Using (C.31), this imposes that

$$C = 0, \quad E = 0 \quad \text{and} \quad \bar{E}_i = 0. \quad (5.92)$$

As the longitudinal one, the flat-slicing gauge is also completely fixed since one can pass from an arbitrary gauge to this one with the transformation (5.58) defined by

$$T = -\frac{C}{\mathcal{H}}, \quad L = -E \quad \text{and} \quad \bar{L}_i = -\bar{E}_i. \quad (5.93)$$

In this gauge, the remaining perturbed quantities are given in terms of the gauge invariant variables by

$$A = X, \quad E' = -\frac{\Psi}{\mathcal{H}}, \quad \delta = \delta^F, \quad \delta P = \delta P^F, \quad v = V \quad \text{and} \quad \bar{v}_i = \bar{W}_i. \quad (5.94)$$

The main advantage of this gauge is, as we shall see, that the conservation equations take the same form as in the Newtonian version of the previous sections.

5.2.3.3 Synchronous or Gauss gauges

Synchronous gauges are defined by the constraints

$$A = 0 \quad \text{and} \quad B_i = 0. \quad (5.95)$$

In such a system of coordinates, only the spatial sections are perturbed, the worldlines with equation $x^i = \text{const.}$ are geodesics orthogonal to constant-time hypersurfaces and the proper time of a comoving observer corresponds to the cosmic time. This is thus a system of coordinates in which every point corresponds to a free-falling observer.

This intuitive physical interpretation had made this gauge very popular. However, it suffers from pathologies, and in particular it is not completely fixed. Indeed, one can still perform coordinate transformations of the form $t \rightarrow t' = f(t)$ or $x^i \rightarrow y^i = f^i(x^j)$, involving only time or space separately and that are compatible with the constraints (5.95). The transformation to pass from an arbitrary gauge to a synchronous gauge is given by

$$T = -\frac{1}{a} \int A a d\eta + \frac{C_T}{a}, \quad L^i = \int (T - B) d\eta + C_L \quad \text{and} \quad \bar{L}^i = - \int \bar{B}^i d\eta + \bar{C}_L^i, \quad (5.96)$$

where C_T and $C_L^i = (C_L, \bar{C}_L^i)$ are four integration functions that can depend on the spatial coordinates x^i . These functions have simple physical interpretations, namely

C_T represents the residual freedom in the choice of the origin of the proper time of each observer, while C_L^t corresponds to an arbitrary choice of the system of coordinates on each constant-time hypersurface. The existence of this arbitrary choice in the definition of the gauge leads to the presence of gauge modes when solving the evolution equations for the perturbations. These gauge modes are not physical and correspond to spurious solutions that may lead (and have led) to confusion. Furthermore, since the gauge is not completely fixed, a variable computed in this gauge cannot always be expressed in terms of gauge invariant quantities.

It is conventional to define the residual metric perturbations in this gauge by

$$h \equiv 6C + 2\Delta E \quad \text{and} \quad \eta = -C. \quad (5.97)$$

5.2.3.4 Comoving gauge

The comoving gauge is a gauge that, as its name indicates, tracks the overall motion of the matter. In this gauge, the velocity of the matter fluid is thus imposed to vanish, i.e.

$$\delta T_i^0 = 0. \quad (5.98)$$

When several fluids are present, a comoving gauge can be defined with respect to any given component (baryons, photons, etc...) or to the total matter. In this text, unless stated otherwise, by ‘comoving gauge’ we really mean ‘comoving gauge with respect to the total matter’.

The condition (5.98) above cannot always be satisfied, as it implies that $v_i + B_i = 0$. It is always possible to find a system of coordinates for which the scalar part of the velocity vanishes (for instance, in a potential flow, Stokes theorem ensures that the flow lines do not intersect and one can thus choose a system of coordinates that follows these lines). This is not the case for the vector part that corresponds to the vorticity and there are current vortices that cannot be suppressed by a change of coordinates. Due to this fact, the following two conditions must be added

$$E = 0 \quad \text{and} \quad \bar{E}_i = 0. \quad (5.99)$$

This gauge has the special feature, reminiscent of one of the flat-slicing gauge property, as we shall see, that the Poisson equation it leads to is identical to that obtained in the Newtonian case of the first sections.

Having shown a few frequently used gauges, let us turn to the actual dynamical description of the perturbation evolution, namely the Einstein equations expanded to first order.

5.2.4 Einstein equations: derivation

Einstein equations for the perturbations take the simple form

$$\delta G_\nu^\mu = \kappa \delta T_\nu^\mu.$$

All required perturbed quantities are gathered in Appendix C. We will treat the tensor modes, followed by the vector ones, which are mathematically simpler, before finishing with the scalar modes.

5.2.4.1 Decomposition of 3-dimensional vector and tensors

Vectors

In Fourier space, we have the splitting

$$V_i = k_i V + \bar{V}_i, \quad \text{with} \quad k^i \bar{V}_i = 0, \quad (5.100)$$

so that \bar{V}^i lives in the subspace perpendicular to k^i . This is a 2-dimensional subspace so that V_i has been split into 1 scalar (V) and two vector modes (\bar{V}_i) that correspond to transverse modes. Let us now consider the base $\{e^1, e^2\}$ of the subspace perpendicular to k^i . By construction, it satisfies the orthonormalization conditions

$$e_i^a k_j \gamma^{ij} = 0, \quad e_i^a e_j^b \gamma^{ij} = \delta^{ab},$$

with $a, b \in \{1, 2\}$. Such a basis is defined up to a rotation about the axis k^i . Now, the vector modes can be decomposed in this basis as

$$\bar{V}_i(k_i, \eta) = \sum_{a=1,2} V_a(\hat{k}_i, \eta) e_i^a(\hat{k}_i). \quad (5.101)$$

This defines the two degrees of freedom, V_a , which depend on \hat{k}^i since the decomposition differs for each wave number. The two basis vectors allow us to define a projection operator onto the subspace perpendicular to k^i as

$$P_{ij} \equiv e_i^1 e_j^1 + e_i^2 e_j^2 = \gamma_{ij} - \hat{k}_i \hat{k}_j, \quad (5.102)$$

where $\hat{k}_i = k_i/k$. It trivially satisfies $P_j^i P_k^j = P_k^i$, $P_j^i k^j = 0$ and $P^{ij} \gamma_{ij} = 2$. It is also the projector on vector modes so that we can always write V_i in the form

$$V_i = (\hat{k}^j V_j) \hat{k}_i + P_i^j V_j, \quad (5.103)$$

thereby making a scalar-vector decomposition, similar to the SVT one for tensors.

Tensors

Analogously, any (3-dimensional) symmetric tensor field, V_{ij} , can be SVT-decomposed as

$$V_{ij} = T \gamma_{ij} + \Delta_{ij} S + 2 D_{(i} \bar{V}_{j)} + 2 \bar{V}_{ij}, \quad (5.104)$$

where we recall that $\Delta_{ij} \equiv D_i D_j - \frac{1}{3} \Delta \gamma_{ij}$ and

$$D_i \bar{V}^i = 0, \quad \bar{V}_i^i = 0 = D_i \bar{V}^{ij}. \quad (5.105)$$

The symmetric tensor \bar{V}_{ij} is, according to our convention, transverse and trace-free. Hence, it has only two independent components and can be decomposed as

$$\bar{V}_{ij}(k_i, \eta) = \sum_{\lambda=+,\times} V_\lambda(k^i, \eta) \epsilon_{ij}^\lambda(\hat{k}_i), \quad (5.106)$$

where the polarization tensors have been defined as

$$\epsilon_{ij}^\lambda = \frac{e_i^1 e_j^1 - e_i^2 e_j^2}{\sqrt{2}} \delta_+^\lambda + \frac{e_i^1 e_j^2 + e_i^2 e_j^1}{\sqrt{2}} \delta_\times^\lambda. \quad (5.107)$$

It can be checked that they are indeed traceless ($\epsilon_{ij}^\lambda \gamma^{ij} = 0$), transverse ($\epsilon_{ij}^\lambda k^i = 0$), and that the two polarizations are perpendicular ($\epsilon_{ij}^\lambda \epsilon_{\mu}^{ij} = \delta_\mu^\lambda$). This defines the two tensor degrees of freedom.

Introducing the projector operator on tensor modes by

$$\Lambda_{ij}^{ab} = P_i^a P_j^b - \frac{1}{2} P_{ij} P^{ab},$$

and the 'trace-extracting' operator

$$T_i^j = \hat{k}_i \hat{k}^j - \frac{1}{3} \delta_i^j,$$

the SVT terms are extracted as follows

$$V_{ij} = \left(\frac{1}{3} V_{ab} \gamma^{ab} \right) \gamma_{ij} + \left(\frac{3}{2} V_{ab} T^{ab} \right) T_{ij} + 2 \hat{k}_{(i} \left[P_{j)}^a \hat{k}^b V_{ab} \right] + \Lambda_{ij}^{ab} V_{ab}. \quad (5.108)$$

In this expression, V_{ij} has been split into 2 scalars (T and S), two vector (\bar{V}_i) and two tensor (\bar{V}_{ij}) modes. Thus, we can always split any vectorial equation such as $V_i = 0$ by projecting it along \hat{k}^i (scalar) and P_i^j (vector), while any tensorial equation of the kind $V_{ij} = 0$ can be projected along γ^{ij} (scalar), T^{ij} (scalar), $P_l^i \hat{k}^j$ (vector) and Λ_{ab}^{ij} (tensor).

Having these decompositions, let us move on to their actual dynamics.

5.2.5 Einstein equations: SVT decomposition

5.2.5.1 Tensor modes

The tensor modes are explicitly gauge invariant by construction. Their evolution equation is obtained by extracting the tensor part of the Einstein equation as

$$\Lambda_{kl}^{ij} \delta G_{ij} = \kappa \Lambda_{kl}^{ij} \delta T_{ij},$$

and they reduce to a single evolution equation

$$\bar{E}_{kl}'' + 2\mathcal{H}\bar{E}'_{kl} + (2K - \Delta)\bar{E}_{kl} = \kappa a^2 P \bar{\pi}_{kl}. \quad (5.109)$$

Up to the curvature term and the damping term due to the expansion, this equation is analogous to (1.132) obtained in Chapter 1 for gravitational waves in a Minkowski space-time.

It can be interesting to rewrite this equation, by introducing the reduced quantity Ω defined in (3.36), in the form

$$\bar{E}_{kl}'' + 2\mathcal{H}\bar{E}'_{kl} + (2K - \Delta)\bar{E}_{kl} = 3\mathcal{H}^2\Omega w\bar{\pi}_{kl}. \quad (5.110)$$

It is also convenient to rewrite the equation for gravitational waves for its two polarizations using the reduced quantity

$$u_\tau = a\bar{E}_\lambda \quad \text{and} \quad \bar{E}_{ij} = \sum_\lambda \bar{E}_\lambda \varepsilon_{ij}^\lambda, \quad (5.111)$$

where ε_{ij}^λ is the polarisation tensor (see Fig. 1.9). Equation (5.109) then takes the form

$$u_\tau'' + \left(2K - \Delta - \frac{a''}{a}\right) u_\tau = \kappa a^3 P \bar{\pi}_\lambda. \quad (5.112)$$

In the case of a Universe with Euclidean spatial sections ($K = 0$) the left-hand side term of this equation is analogous to that of a field evolving in a potential, here time dependent, given by a''/a . Note that this equation also has the same mathematical form as a time-independent one-dimensional Schrödinger equation for a particle in a potential, the latter being $d^2\psi(x)/dx + [E - V(x)]\psi(x) = 0$. This formulation will come in handy when studying gravitational waves during inflation (see Chapter 8).

5.2.5.2 Vector modes

There are two Einstein equations for the vector modes that follow, respectively, from the components $(0i)$ and (ij) . Using the results of Appendix C and the expressions (5.75) and (5.76) for the perturbed energy-momentum tensor, we readily obtain the two equations

$$(\Delta + 2K)\bar{\Phi}_i = -2\kappa\rho a^2(1+w)\bar{V}_i, \quad (5.113)$$

$$\bar{\Phi}'_i + 2\mathcal{H}\bar{\Phi}_i = \kappa P a^2 \bar{\pi}_i. \quad (5.114)$$

As for the case of (5.110), it can also be interesting to rewrite these two equations, by introducing the reduced quantity Ω , defined in (3.36), in the form

$$(\Delta + 2K)\bar{\Phi}_i = -6\mathcal{H}^2\Omega(1+w)\bar{V}_i, \quad (5.115)$$

$$\bar{\Phi}'_i + 2\mathcal{H}\bar{\Phi}_i = 3\mathcal{H}^2\Omega w\bar{\pi}_i. \quad (5.116)$$

5.2.5.3 Scalar modes

To derive the Einstein equations for the scalar modes, we will consider the following four combinations:

- $\delta G_0^0 + 3\mathcal{H}D_i^{-1}\delta G_i^0$,
- the traceless part of δG_j^i ,
- δG_i^0 ,

$$-\delta G_i^i + 3c_s^2 \delta G_0^0.$$

First, we obtain two constraint equations,

$$(\Delta + 3K)\Psi = \frac{\kappa}{2} a^2 \rho \delta^c, \quad (5.117)$$

$$\Psi - \Phi = \kappa a^2 P \bar{\pi}. \quad (5.118)$$

The first equation, as announced above, takes the classical form of the Poisson equation when expressed in terms of δ^c , using (5.87). The second equation tells us that the two gravitational potentials are equal if the scalar component of the anisotropic stress tensor vanishes (most frequent case).

The other two equations are equations of evolution given by

$$\Psi' + \mathcal{H}\Phi = -\frac{\kappa}{2} a^2 \rho (1+w)V, \quad (5.119)$$

$$\begin{aligned} \Psi'' + 3\mathcal{H}(1+c_s^2)\Psi' + [\mathcal{H}' + (\mathcal{H}^2 - K)(1+3c_s^2)]\Psi - c_s^2 \Delta \Psi = \\ -(\mathcal{H}^2 + 2\mathcal{H}' + K) \left[\frac{1}{2}\Gamma + (3\mathcal{H}^2 + 2\mathcal{H}')\bar{\pi} + \mathcal{H}\bar{\pi}' + \frac{1}{3}\Delta\bar{\pi} \right] \\ - 9c_s^2 \mathcal{H}^2 (\mathcal{H}^2 + K) \bar{\pi}, \end{aligned} \quad (5.120)$$

where we have used, in the second equation, (5.118) to replace Φ by its value in terms of Ψ and $\bar{\pi}$.

5.2.5.4 Equivalent forms

It can be useful to give some equivalent forms of the Einstein equations for the scalar perturbations, in particular using other gauge invariant density contrasts.

The Poisson equation (5.117) takes the following two forms depending on whether δ^N or δ^F is used

$$(\Delta + 3K)\Psi = \frac{\kappa}{2} a^2 \rho [\delta^N - 3\mathcal{H}(1+w)V], \quad (5.121)$$

$$\Delta\Psi - 3\mathcal{H}^2 X = \frac{\kappa}{2} a^2 \rho \delta^F, \quad (5.122)$$

while (5.120) can be expressed in terms of the variable X alone as

$$\mathcal{H}X' + (\mathcal{H}^2 + 2\mathcal{H}')X = \frac{\kappa}{2} a^2 \rho \left(w\Gamma + c_s^2 \delta^F + \frac{2}{3}w\Delta\bar{\pi} \right). \quad (5.123)$$

5.2.6 Perturbed conservation equation for a fluid

The fluid conservation equations take the simple form

$$\delta(\nabla_\mu T^\mu_\nu) = 0.$$

Again, all the required perturbed quantities are gathered in Appendix C. We treat the vector modes first, and then the scalar ones. Note that, since fluids are described by scalar and vector quantities only, there is no tensor equation.

5.2.6.1 Vector modes

There is a unique conservation equation for the vector modes and it takes the form

$$\bar{V}'_i + \mathcal{H}(1 - 3c_s^2)\bar{V}_i = -\frac{1}{2}\frac{w}{1+w}(\Delta + 2K)\bar{\pi}_i. \quad (5.124)$$

This form is sufficiently simple that one can solve it explicitly in most of the interesting cases.

5.2.6.2 Scalar modes

For the scalar modes, the 4-dimensional conservation equations will give us two equations analogous to the continuity equation and to the Euler equation, obtained, respectively, from the time-like and space-like component.

Working in the Newtonian gauge, one obtains

$$\delta^N' + 3\mathcal{H}(c_s^2 - w)\delta^N = -(1 + w)(\Delta V - 3\Psi') - 3\mathcal{H}w\Gamma, \quad (5.125)$$

$$V' + \mathcal{H}(1 - 3c_s^2)V = -\Phi - \frac{c_s^2}{1+w}\delta^N - \frac{w}{1+w}\left[\Gamma + \frac{2}{3}(\Delta + 3K)\bar{\pi}\right]. \quad (5.126)$$

The conservation equation (5.125) can be rewritten in the form

$$\left(\frac{\delta^N}{1+w}\right)' = -(\Delta V - 3\Psi') - 3\mathcal{H}\frac{w}{1+w}\Gamma, \quad (5.127)$$

if we use the expression (C.19) to re-express w' . It is also useful to give the expression for these two equations in terms of the density contrasts δ^F and δ^C , namely

$$\delta^F' + 3\mathcal{H}(c_s^2 - w)\delta^F = -(1 + w)\Delta V - 3\mathcal{H}w\Gamma, \quad (5.128)$$

$$V' + \mathcal{H}(1 - 3c_s^2)V = -\Phi - 3c_s^2\Psi - \frac{c_s^2}{1+w}\delta^F - \frac{w}{1+w}\left[\Gamma + \frac{2}{3}(\Delta + 3K)\bar{\pi}\right], \quad (5.129)$$

$$V' + \mathcal{H}V = -\Phi - \frac{c_s^2}{1+w}\delta^C - \frac{w}{1+w}\left[\Gamma + \frac{2}{3}(\Delta + 3K)\bar{\pi}\right]. \quad (5.130)$$

5.2.7 Interpretation of the perturbation equations

At first glance, the conservation equations are analogous to those obtained in the Newtonian approximation. In particular, as announced before, we recover (5.1) and (5.2) if we place ourselves in a static space ($\mathcal{H} = 0$) with non-relativistic matter ($w = c_s^2 = 0$ and $\bar{\pi} = 0$) for adiabatic perturbations ($\Gamma = 0$) if δ is identified as δ^F . However, there are some important differences we ought to stress.

1. There are vector and tensor perturbations in addition to the scalar perturbations.

2. There are two potentials, Φ and Ψ , to describe the scalar perturbations of the metric. These should tend to the same value on small scales to be compatible with the Newtonian limit, in which the anisotropic stress tensor becomes negligible. It is in the comoving gauge that the Poisson equation is the closest to its Newtonian form but note that it involves the potential Ψ rather than Φ .
3. On large scales, the Poisson equation differs from its Newtonian analogue and the different density contrasts can have very different values.
4. The energy density conservation equation has an additional factor $(1+w)$ as it is the energy flux rather than the matter flux that comes in at the relativistic level.
5. The Euler equation has an additional damping term $-\mathcal{H}(1-3c_s^2)$. Typically, this term will have the effect of reducing the velocity as a^{-1} for any non-relativistic matter: the velocity field tends to align with the Hubble flow. For ultrarelativistic matter, this term vanishes, which can be understood by going back to geodesics and noticing that the null geodesics of two conformal space-times are identical. As will be seen later, a proper description of radiation should be done in terms of a kinetic theory. Velocity actually represents the dipolar term of the radiation distribution (see Chapter 6).

5.3 Evolution

The perturbation equations derived in the previous sections can be solved in various simple cases. In general, it is useful to expand the perturbations into Fourier modes (see Appendix B) and to study the evolution of each mode as a function of time. Indeed, since the equations of evolution involve spatial derivatives only through the Laplacian, they will reduce to linear ordinary differential equations in Fourier space.

In many cases, we will distinguish two regimes that we introduce now

| | | |
|---------------------|-------------------|--|
| $k \ll \mathcal{H}$ | super-Hubble mode | wavelength larger than Hubble radius, |
| $k \gg \mathcal{H}$ | sub-Hubble mode | wavelength smaller than Hubble radius. |

Note that in the literature, the super- and sub-Hubble modes are often, wrongly (and confusingly), called sub- and superhorizon. This (historically based) misuse of language can be understood as in the standard Big-Bang model presented in Chapter 4 the Hubble radius coincides, up to an irrelevant numerical factor, with the particle horizon radius. However, this is not always the case, in particular if there has been a period of inflation, which precisely has, among its aims, the purpose of solving the horizon problem (see Chapter 8)! The fact that the spatial gradients are negligible in the evolution equations of the perturbations is not at all related to causality but to the comparison between the damping term due to the expansion of the Universe, \mathcal{H}^{-1} and the characteristic time of oscillation of mode k , given by $\sim k^{-1}$.

Note that the behaviour of the growing mode on super-Hubble scales cannot be inferred intuitively from causality arguments, contrary to what is sometimes stated

when claiming that on superhorizon scales the perturbations are ‘frozen’ (see, e.g., Ref. [34] for a detailed discussion).

5.3.1 Vector and tensor modes

5.3.1.1 Vector modes

As long as $\bar{\pi}_i$ vanishes or is negligible, (5.114) implies that

$$\bar{\Phi}_i \propto a^{-2}, \quad (5.131)$$

and the Euler equation implies that

$$\bar{V}^i \propto a^{-(1-3\epsilon_s^2)}. \quad (5.132)$$

So that the vector metric perturbations are washed out by the cosmological evolution. The same is true for the vorticity of the cosmic fluid if the latter is non-relativistic. Thus, the vector modes are of little interest in the formation of the large-scale structures, even if they can play a role, for example, if magnetic fields or topological defects are present.

5.3.1.2 Tensor modes

Let us assume that $\bar{\pi}_{ij}$ vanishes or is negligible and let us write (5.109) in Fourier space for a Universe with Euclidean spatial sections. Assuming that the scale factor behaves as $a \propto \eta^\nu$ then

$$\frac{d^2 \bar{E}_{ij}}{dx^2} + \frac{2\nu}{x} \frac{d\bar{E}_{ij}}{dx} + \bar{E}_{ij} = 0, \quad (5.133)$$

with $x \equiv k\eta$. The solution of this equation can be expressed in terms of Bessel functions as

$$\bar{E}_{ij} = x^{1/2-\nu} \left[A_{ij} J_{\nu-\frac{1}{2}}(x) + B_{ij} N_{\nu-\frac{1}{2}}(x) \right], \quad (5.134)$$

where A_{ij} and B_{ij} are two transverse and traceless constant tensors. Keeping only the mode that does not diverge when $x \rightarrow 0$ (the divergent mode corresponding to a decaying mode when η increases indeed becomes negligible) we obtain

$$\bar{E}_{ij} = x^{1/2-\nu} J_{\nu-\frac{1}{2}}(x) A_{ij} = x^{1-\nu} j_{\nu-1}(x) A_{ij}. \quad (5.135)$$

In particular, we obtain, respectively, in the matter- and radiation-dominated epochs

| | Matter | Radiation |
|----------------------|---------------------|--------------|
| $a(\eta)$ | η^2 | η |
| $\bar{E}_{ij}(\eta)$ | $3j_1(k\eta)/k\eta$ | $j_0(k\eta)$ |

\bar{E}_{ij} is constant as long as $k\eta < 1$, i.e. as long as the wavelength of the considered mode is larger than the Hubble radius. When the wavelength becomes shorter, the mode develops a damped oscillatory behaviour (see Fig. 5.3).

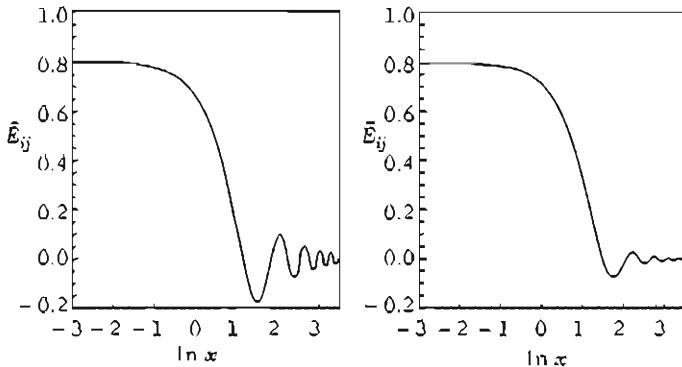


Fig. 5.3 Evolution of gravitational waves in a radiation-dominated Universe, i.e. $w = \frac{1}{3}$, $\nu = 1$, (left) and pressureless matter, $w = 0$, $\nu = 2$ (right) in terms of $\ln x = \ln k\eta$. The mode is constant as long as the wavelength is super-Hubble ($x < 1$) and develops a damped oscillatory behaviour as soon as it becomes sub-Hubble ($x > 1$).

5.3.1.3 Tensor modes in a Universe dominated by a mixture of matter and radiation
Let us now assume that the Universe is dominated by a mixture of pressureless matter ($w = 0, \nu = 2$) and radiation ($w = \frac{1}{3}, \nu = 1$). Introducing the quantity

$$y \equiv \frac{a}{a_{eq}}, \quad (5.136)$$

where a_{eq} is the value of the scale factor when radiation and matter have the same density [$\rho_m(a_{eq}) = \rho_r(a_{eq})$], we deduce that $y = \rho_m/\rho_r$ and that the equation of state of the mixture of the two fluids is

$$w = \frac{1}{3} \frac{1}{1+y}. \quad (5.137)$$

The evolution (5.109) of the gravitational waves can then be rewritten as

$$\frac{d^2 \bar{E}_{ij}}{dy^2} + \frac{4+5y}{2y(1+y)} \frac{d\bar{E}_{ij}}{dy} + \left(\frac{k}{k_{eq}}\right)^2 \frac{2}{1+y} \bar{E}_{ij} = 0, \quad (5.138)$$

where we have switched to the variable y : to obtain this equation, we have used the fact that the derivatives with respect to y are given by $\bar{E}'_{ij} = y' d\bar{E}_{ij}/dy$, and that the Hubble parameter is given by $\mathcal{H} = y'/y$. The Friedmann equation, with this variable, takes the form

$$\mathcal{H}^2 = \mathcal{H}_{eq}^2 \frac{1+y}{2y^2}, \quad (5.139)$$

where \mathcal{H}_{eq} is the value of \mathcal{H} at a_{eq} . We have also introduced k_{eq} , the value of the wave number corresponding to the mode entering the Hubble radius at equality, defined by

$$k_{eq} = \mathcal{H}_{eq} = a_{eq} H_{eq}. \quad (5.140)$$

Equation (5.138) can be integrated numerically and the solutions for different values of k/k_{eq} are represented in Fig. 5.4. Note that the earlier the mode enters the Hubble radius, the more damped it is.

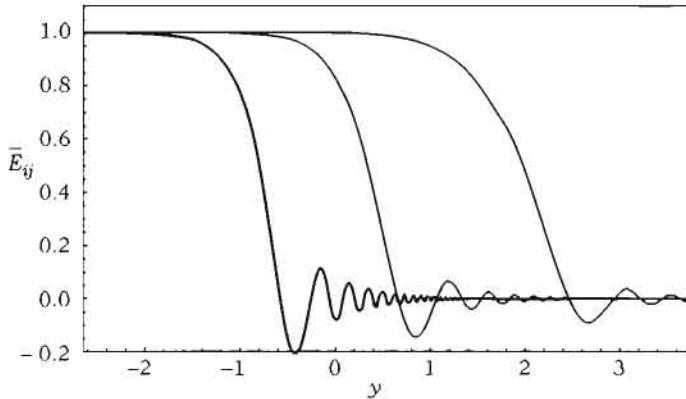


Fig. 5.4 Evolution of gravitational waves in a Universe dominated by a mixture of matter and radiation in terms of $y = a/a_{eq}$ ($y = 1$ at equality) for the modes of value $k/k_{eq} = 10, 1$ and 0.1 .

5.3.1.4 General solutions for long-wavelength modes

In a flat Universe ($K = 0$), the form (5.112) of the gravitational waves evolution equation admits the (formal) integral solution

$$u_T(\eta, k) = a(\eta) \left[A_1(k) + A_2(k) \int^{\eta} \frac{d\bar{\eta}}{a^2(\bar{\eta})} + k^2 \int^{\eta} \frac{d\bar{\eta}}{a^2(\bar{\eta})} \int^{\bar{\eta}} d\bar{\eta} a^2(\bar{\eta}) u_T(\bar{\eta}, k) \right]. \quad (5.141)$$

For long wavelength modes, that is satisfying $k \ll a''/a$, it reduces to

$$u_T(\eta, k) = a(\eta) A_1(k) + A_2(k) a(\eta) \int^{\eta} \frac{d\bar{\eta}}{a^2(\bar{\eta})} + \mathcal{O}\left(\frac{k^2 a}{a''}\right), \quad (5.142)$$

where the two integration constants A_1 and A_2 may a priori depend on k . Note, in the analogy with the Schrödinger equation, that this is simply the first term of a Born expansion for describing the deflection of a particle in a potential.

5.3.2 Evolution of the gravitational potential

We are interested in the evolution of the gravitational potential, so let us rewrite its evolution equation in the form

$$\Phi'' + 3\mathcal{H}(1 + c_s^2)\Phi' + [2\mathcal{H}' + (\mathcal{H}^2 - K)(1 + 3c_s^2)]\Phi - c_s^2\Delta\Phi = \frac{\kappa}{2}a^2P\Gamma, \quad (5.143)$$

when we neglect the contributions from the anisotropic stress ($\bar{\pi} = 0$).

5.3.2.1 Another formulation of the evolution equation

A reduced form of this evolution equation can be obtained. First, we introduce the auxiliary function

$$\theta = \frac{1}{a} \left(\frac{\rho}{\rho + P} \right)^{1/2} \left(1 - \frac{3K}{\kappa \rho a^2} \right)^{1/2}, \quad (5.144)$$

which can be rewritten as

$$\theta = \frac{\mathcal{H}}{a} \left[\frac{2}{3} (\mathcal{H}^2 - \mathcal{H}' + K) \right]^{-1/2}. \quad (5.145)$$

We then define the auxiliary variable

$$u_s = \frac{2}{3} \frac{a^2 \theta}{\mathcal{H}} \Phi, \quad (5.146)$$

thanks to which, and after a tedious, but straightforward, computation, one can show that (5.120) eventually takes the form

$$u_s'' - \left(\frac{\theta''}{\theta} + c_s^2 \Delta \right) u_s = \frac{\kappa}{3} \frac{\theta}{\mathcal{H}} a^4 P \Gamma.$$

(5.147)

which shares many features with the form (5.112) and can thus be analysed using similar tools.

5.3.2.2 Integral solution for long-wavelength modes

The homogeneous part of this equation can be integrated for any long-wavelength mode by noticing that $u = \theta$ is a solution. In Fourier space, we obtain

$$u_s(k, \eta) = A(k)\theta(\eta) + B(k)\theta(\eta) \int^\eta \frac{d\eta'}{\theta^2(\eta')} + \mathcal{O}\left(\frac{k^2 c_s^2 \theta}{\theta''}\right). \quad (5.148)$$

This solution is valid for all modes in the matter era as long as the entropy perturbation Γ is negligible. As discussed above, it can be compared with (5.142) by replacing a by θ .

5.3.2.3 First integral

Equation (5.143) admits a first integral for the adiabatic evolution of super-Hubble modes [35]. Indeed, by introducing the quantity [36]

$$\zeta \equiv \Phi + \frac{2}{3\mathcal{H}} \frac{\Phi' + \mathcal{H}\Phi}{1+w}, \quad (5.149)$$

(5.143) takes the form

$$\zeta' = \frac{2}{3} \frac{\mathcal{H}}{1+w} \left\{ \frac{K}{\mathcal{H}^2} \left[\frac{1+3w}{2\mathcal{H}} \Phi' + 3(c_s^2 - w) \Phi \right] + \frac{\kappa}{2} \frac{a^2}{\mathcal{H}^2} P \Gamma - c_s^2 \left(\frac{k}{\mathcal{H}} \right)^2 \Phi \right\}, \quad (5.150)$$

where we have used (C.19) and (C.22) to evaluate w' and \mathcal{H}' . We infer that as long as the curvature is negligible, i.e. in the regime for which $K/\mathcal{H}^2 \ll 1$, the adiabatic

evolution ($\Gamma = 0$) of the super-Hubble modes ($k/\mathcal{H} \ll 1$) has ζ as first integral, namely that the approximation

$$\zeta' = 0 \quad (5.151)$$

is valid under these assumptions. Note that if the evolution is not adiabatic, then (5.150) takes the form

$$\zeta' = \frac{\mathcal{H}}{\rho + P} \delta P_{\text{nad}}, \quad (5.152)$$

so the conservation of ζ is not ensured if entropy perturbations are produced. The quantity ζ turns out to have a very simple geometrical interpretation: considering the expression (C.31) for the spatial curvature, we notice that ζ is exactly this curvature in the comoving gauge.

Let us also introduce the curvature perturbation in the flat-slicing gauge as

$$\zeta_{\text{BSR}} \equiv -C + \frac{1}{3} \frac{\delta\rho}{\rho + P} = -C + \frac{1}{3} \frac{\delta}{1+w}. \quad (5.153)$$

The variable ζ_{BSR} , being gauge invariant, can be expressed in terms of the gravitational potential as

$$\zeta_{\text{BSR}} = \Phi - \frac{2}{3(1+w)(\mathcal{H}^2 + K)} \left\{ \Phi' + \left[1 - \frac{K}{\mathcal{H}^2} + \frac{1}{3} \left(\frac{k}{\mathcal{H}} \right)^2 \right] \mathcal{H}\Phi \right\}. \quad (5.154)$$

We infer that ζ_{BSR} therefore satisfies

$$\zeta'_{\text{BSR}} = -\frac{2}{3(1+w)(\mathcal{H}^2 + K)} \left[\frac{1}{3} \left(\frac{k}{\mathcal{H}} \right)^2 (\Phi' + \mathcal{H}\Phi) + \frac{\kappa a^2}{2 \mathcal{H}^2} P\Gamma \right]. \quad (5.155)$$

Note that unlike (5.150), the right-hand side term of this equation does not involve the spatial curvature K . Thus, ζ_{BSR} remains constant in time for the adiabatic ($\Gamma = 0$) and super-Hubble ($k/\mathcal{H} \ll 1$) modes provided that $\Phi' + \mathcal{H}\Phi$ does not diverge when $k/\mathcal{H} \rightarrow 0$, which is the case if the decaying mode is negligible.

It is easy to see that both quantities differ only on sub-Hubble scales and that they are related by

$$\zeta = \zeta_{\text{BSR}} - \frac{1}{3} \frac{\Delta\Phi}{\mathcal{H}' - \mathcal{H}^2}. \quad (5.156)$$

It should be noted at this point that the conservation of ζ_{BSR} , although derived in the Einstein equations framework, does not in fact depend on the underlying gravity theory, but only on $\nabla_\mu T^{\mu\nu} = 0$, i.e. the fact that the stress tensor is conserved [37]. This assumption is much weaker and renders this variable useful in wider frameworks.

5.3.2.4 Effect of a variation in the cosmic fluid equation of state

For illustrative purposes, let us consider the evolution of the gravitational potential during a change in the equation of state. We thus assume that w varies from w_1 for $\eta < \eta_*$ (era 1) to w_2 for $\eta > \eta_*$ (era 2) and we assume that $K = 0$ and $\Gamma = 0$.

In each phase, the scale factor evolves as $a \propto t^{p_i} \propto \eta^{n_i}$, with $n_i = p_i/(1-p_i)$ and $1+w_i = 2/(3p_i)$.

The general solution (5.148) for long-wavelength modes implies

$$\Phi = \frac{\mathcal{H}}{a^2} \left[A_- + A_+ \int a^2(1+w)d\eta \right], \quad (5.157)$$

where A_+ and A_- are two constants that characterize the growing and decaying modes. If we assume that the decaying mode in the era I has had enough time to become negligible, then, before the transition, the only mode that remains is the constant one, namely $\Phi(\eta < \eta_*) \sim A_+ (1+w_1) p_1/(p_1+1) = 2A_+/[3(1+p_1)]$, while at a large enough time after the transition, when the decaying term is also negligible, one similarly ends up with $\Phi(\eta \gg \eta_*) \sim 2A_+/[3(1+p_2)]$. We then have

$$\frac{\Phi(\eta \gg \eta_*)}{\Phi(\eta < \eta_*)} = \frac{1+p_1}{1+p_2}. \quad (5.158)$$

In the special case of a transition between a radiation-dominated era ($p_1 = \frac{1}{2}$) and a matter-dominated era ($p_2 = \frac{2}{3}$), this leads to

$$\frac{\Phi(\eta \gg \eta_*)}{\Phi(\eta < \eta_*)} = \frac{9}{10}, \quad (5.159)$$

for long-wavelength modes.

Note that one could have reached the same conclusion from the fact that ζ is constant during the evolution, since from (5.149), $\zeta \sim \{1+2/[3(1+w)]\}\Phi$ for the growing mode, during each era. As it turns out, in this particular case, the relation (5.159) can also be obtained by using the analytical solution for the scale factor [27].

5.3.3 Scalar modes in the adiabatic regime

We assume in the following examples that the influence of the anisotropic stress tensor and of the entropy perturbation are negligible ($\bar{\pi} = 0$ and $\Gamma = 0$). In particular, this implies that the two gravitational potentials are equal, $\Phi = \Psi$. Thus, this chapter will mainly focus on the study of the evolution equation (5.143) with $\Gamma = 0$. Note that these assumptions are valid in most of the evolution of the Universe.

5.3.3.1 Universe dominated by a fluid with equation of state $w \neq 0$

Let us consider a flat Universe ($K = 0$) dominated by a fluid with constant equation of state $w \neq 0$. In particular, this implies that $c_s^2 = w$. The scale factor of such a Universe evolves as $a \propto \eta^\nu$ with $\nu = 2/(1+3w)$, so that in Fourier space, the evolution equation (5.143) takes the form

$$\frac{d^2 f}{dx^2} + \frac{2}{x} \frac{df}{dx} + \left[w - \frac{\nu(\nu+1)}{x^2} \right] f = 0, \quad (5.160)$$

after introducing the function $f = x^\nu \Phi$ and $x = k\eta$. The general solution of this equation is known and yields the gravitational potential. The latter can be expressed in terms of Bessel functions as

$$\Phi = \frac{-3\nu^2}{2} x^{-\nu} [A j_\nu(c_s x) + B n_\nu(c_s x)] \equiv \frac{-3\nu^2}{2} x^{-\nu} Z_\nu(c_s x), \quad (5.161)$$

where j_ν and n_ν are spherical Bessel functions defined in Appendix B, and where the second equality defines Z_ν . The Poisson equation (5.117) allows us to deduce that the density contrast is given by

$$\delta^C = x^{2-\nu} Z_\nu(c_s x), \quad (5.162)$$

and (5.119) then implies that

$$kV = -\frac{3}{4} x^{1-\nu} \left[Z_\nu(c_s x) - \frac{c_s x}{\nu+1} Z_{\nu-1}(c_s x) \right]. \quad (5.163)$$

Using the asymptotic behaviour of the Bessel functions (see Appendix B), we obtain the following behaviour in the two limiting regimes

| | $c_s x \ll 1$ | $c_s x \gg 1$ |
|-------------|---|--|
| Φ | $\Phi_+ + \Phi_- x^{-1-2\nu}$ | $\Phi_+ x^{-(1+\nu)} \cos(c_s x - \alpha_\nu)$ |
| $-\delta^C$ | $2(\Phi_+ x^2 - \Phi_- x^{1-2\nu}) / (3\nu^2)$ | $2\Phi_+ x^{1-\nu} \cos(c_s x - \alpha_\nu) / (3\nu^2)$ |
| $-kV$ | $2[\Phi_+ x - \Phi_- (1 + \nu^{-1}) x^{-2\nu}] / [3\nu(1+w)]$ | $2\Phi_+ x^{1-\nu} \cos(c_s x - \alpha_\nu) / [3\nu(1+w)]$ |

The phase α_ν is given by $\alpha_\nu = \pi(\nu+1)/2$. For wavelengths much shorter than the acoustic radius ($c_s x \gg 1$), δ^C behaves as a sound wave that is damped if $-\frac{1}{2} < w < \frac{1}{3}$. Radiation ($w = \frac{1}{3}$) is the limiting case where the amplitude of the density perturbation remains constant (Fig. 5.5). Outside the acoustic radius ($c_s x \ll 1$), the growing mode of the gravitational potential remains constant.

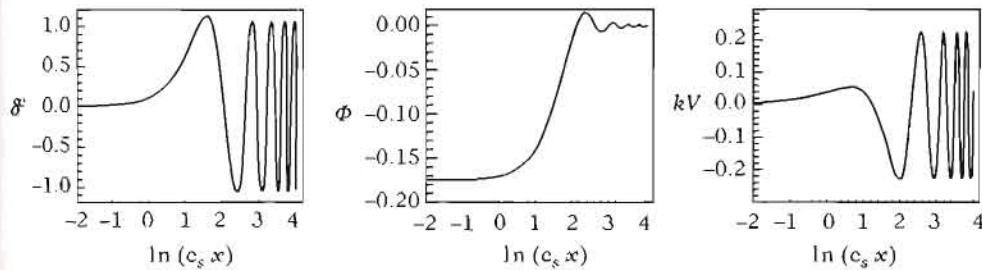


Fig. 5.5 Evolution of the density contrast (left), of the gravitational potential (middle) and of the velocity perturbation (right) for the scalar modes of a radiation fluid ($w = c_s^2 = \frac{1}{3}$) in a radiation-dominated Universe ($\nu = 1$) in terms of $c_s x = c_s k \eta$. The mode is constant as long as the wavelength is larger than the sound radius ($c_s x < 1$) and it starts a damped oscillatory behaviour as soon as it is smaller ($c_s x > 1$).

We also deduce from these calculations that during the radiation era, the super-Hubble density fluctuations of the radiation fluid grow as the square of the scale factor,

$$\delta_r^C \propto a^2. \quad (5.164)$$

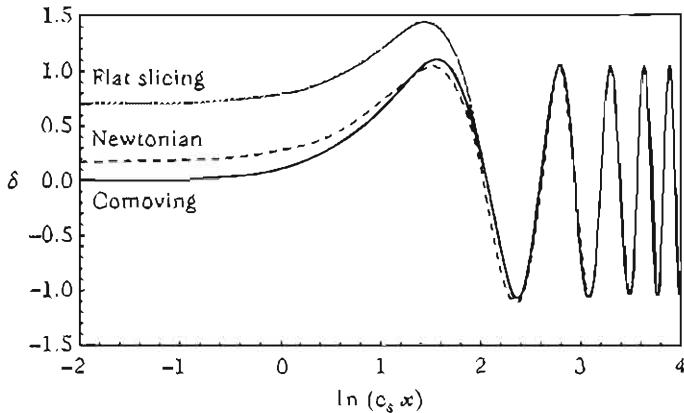


Fig. 5.6 Evolution of the density contrasts δ^C (solid line), δ^N (dashed) and δ^F (dotted) in terms of $c_s x = c_s k\eta$. As long as the wavelength is larger than the sound radius ($c_s x < 1$) there is a difference between these three quantities that converge rapidly as soon as they become smaller than the sound radius ($c_s x > 1$). This illustrates the effect of the gauge choice.

5.3.3.2 Matter fluid in a matter-dominated Universe

In the case of a dust-dominated Universe ($w = c_s^2 = 0$ and $\nu = 2$), and using the same notations as in the previous section, (5.160) takes the form,

$$\frac{d^2 f}{dx^2} + \frac{2}{x} \frac{df}{dx} - \frac{6}{x^2} f = 0, \quad (5.165)$$

whose general solution provides the potential, density contrast and velocity, namely

$$\Phi = \Phi_+ + \Phi_- x^{-5}, \quad \delta^C = -\frac{1}{6} x^2 \Phi \quad \text{and} \quad -kV = \frac{1}{3} \Phi_+ x - \frac{1}{2} \Phi_- x^{-4}. \quad (5.166)$$

We conclude (since $a \propto \eta^2 \propto x^2$) that the density contrast of matter grows as

$$\delta_m^C \propto a, \quad (5.167)$$

independently of the wavelength, which is the conclusion we had reached from the Newtonian analysis for sub-Hubble modes [see (5.22)]. The constancy of the gravitational potential in a matter-dominated Universe and the fact that the density contrast cannot grow faster than the scale factor led Landau in 1946 to the conclusion that the gravitational instability in an expanding space-time could not explain the large-scale structures observed in the Universe [38]. Indeed, if the initial fluctuations are simple thermal fluctuations, they can only grow, according to (5.167), by a mere factor of 10^4 during the matter-dominated era, which is far too small to explain the existence of the present fluctuations. We thus need to find a mechanism capable of generating perturbations of the order of 10^{-4} at the beginning of the matter-dominated era in order for those to have enough time to grow non-linear by the current epoch.

5.3.3.3 Effect of curvature

Curvature, behaving as a^{-2} , can only dominate at late times, i.e. at a time at which radiation is negligible, i.e. during the matter-dominated epoch we are considering here. The solution for the evolution of the scale factor in such a matter and curvature Universe was obtained before, in Section 3.2.1. If we only consider the matter-curvature transition (i.e. the case $\alpha = \frac{1}{2}$), we can integrate (5.143) analytically to obtain

$$\Phi(k, \eta) = A(k) \frac{\sinh \eta}{\sinh^6(\eta/2)} + B(k) \frac{\sinh \eta (\sinh \eta - 3\eta) + 8 \sinh^2(\eta/2)}{\sinh^6(\eta/2)}, \quad (5.168)$$

where we have assumed $K = -1$. Note that as soon as the curvature dominates ($\eta > 1$) then $\Phi \sim B(k)e^{-\eta}$ and the density contrast can be found using (5.117), namely

$$\delta_m^c = -8k^2 \frac{|1 - \Omega_0|}{\Omega_0} \sinh^2\left(\frac{\eta}{2}\right) \Phi \sim -8k^2 B(k) \frac{|1 - \Omega_0|}{\Omega_0}. \quad (5.169)$$

Thus, as soon as the curvature dominates, the gravitational potential exponentially decays and the density contrast freezes.

For a positive curvature, i.e. $K = +1$, the solution of (5.143) changes to

$$\Phi(k, \eta) = A(k) \frac{\sin \eta}{\sin^6(\eta/2)} + B(k) \frac{\sin \eta (\sin \eta - 3\eta) + 8 \sin^2(\eta/2)}{\sin^6(\eta/2)}, \quad (5.170)$$

and we now have two distinct behaviours. When $\eta \rightarrow \pi - \varepsilon$, one approaches the maximum of the expansion cycle, \mathcal{H} becomes vanishingly small, and $\Phi \sim 8B(k)$. The Universe then enters a subsequent contracting phase, during which it is worth noting that both modes diverge: the contraction amplifies the gravitational collapse. Just before the big crunch at $\eta = 2\pi - \varepsilon$, the potential is $\Phi \sim 2(6\pi B - A)/\sin^5(\eta/2)$ and the density contrast evolves as

$$\delta_m^c \sim -2k^2 \frac{6\pi B(k) - A(k)}{\sin^3(\eta/2)} \frac{|1 - \Omega_0|}{\Omega_0}. \quad (5.171)$$

5.3.3.4 Effect of the cosmological constant

A cosmological constant can also only dominate at late times, i.e. at a time where radiation is negligible. Introducing $y_\Lambda = a/a_\Lambda = (\Lambda/\kappa\rho)^{1/3}$, one readily finds that the transition between the matter-dominated era and the cosmological-constant dominated era occurs at redshift

$$1 + z_\Lambda = \left(\frac{\Omega_{\Lambda 0}}{\Omega_{m0}} \right)^{1/3},$$

and the Friedmann equation takes the form

$$\mathcal{H}^2 = \frac{\mathcal{H}_\Lambda^2}{2} \left(\frac{1}{y_\Lambda} + y_\Lambda^2 \right).$$

Equation (5.143) then takes the form

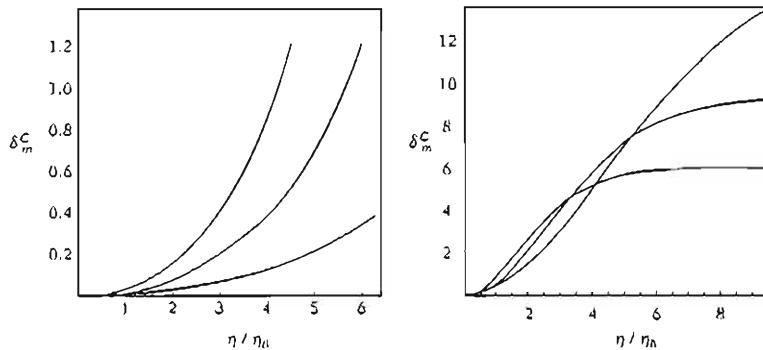


Fig. 5.7 Evolution of the density contrast δ_m^C in a Universe with spherical spatial section (left) for $\Omega_0 = 1.2, 1.3, 1.4$ and a Universe with hyperbolic spatial sections (right) for $\Omega_0 = 0.2, 0.3, 0.4$.

$$\ddot{\Phi} + \frac{1}{2y_\Lambda} \frac{7 + 10y_\Lambda^3}{1 + y_\Lambda^3} \dot{\Phi} + \frac{3y_\Lambda}{y_\Lambda^3 + 1} \Phi = 0. \quad (5.172)$$

The density contrast is then given by

$$\delta_m^C = -\frac{2}{3} \left(\frac{k}{H_\Lambda} \right)^2 y_\Lambda \Phi. \quad (5.173)$$

As soon as $y_\Lambda \gg 1$, the potential decays as $1/y_\Lambda$ and the density contrast freezes to a constant value (see Fig. 5.8). In such a Universe the large-scale structure formation stops at a redshift of the order of z_Λ .

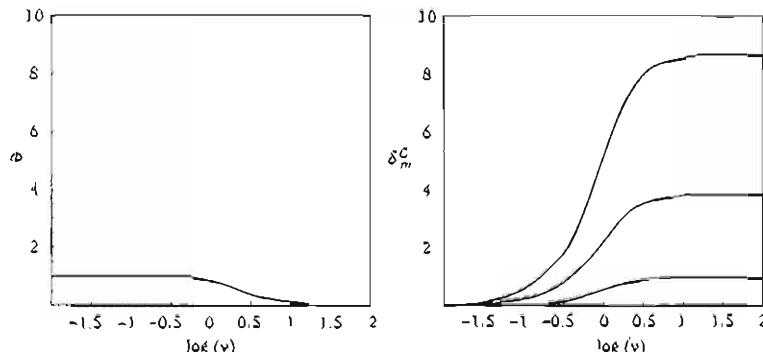


Fig. 5.8 Evolution of the gravitational potential (left) and of the density contrast (right) in a Universe dominated by a cosmological constant. On the right, we have shown the three modes $k/H_\Lambda = 1, 2, 3$.

5.3.3.5 Case of a Universe dominated by a mixture of matter and radiation

Let us now consider the case of a flat Universe ($K = 0$) dominated by a mixture of matter ($w = 0$) and radiation ($w = \frac{1}{3}$). We again introduce the variable y , the scale factor (5.136) normalized at equality and we have

$$w = \frac{1}{3}(1+y)^{-1} \quad \text{and} \quad c_s^2 = \frac{1}{3} \left(1 + \frac{3}{4}y\right)^{-1}. \quad (5.174)$$

Using $\Phi' = \mathcal{H}y\Phi/dy$, $\Phi'' = \mathcal{H}^2y^2d^2\Phi/dy^2 + y''d\Phi/dy$ and the form (5.139) of the Friedmann equation, we can rewrite the evolution (5.143) as

$$\begin{aligned} \frac{d^2\Phi}{dy^2} + \frac{1}{2y} \left(7 - \frac{1}{1+y} + \frac{8}{4+3y}\right) \frac{d\Phi}{dy} + \frac{1}{y(1+y)(4+3y)}\Phi \\ + \frac{8}{3} \frac{1}{4+3y} \frac{1}{1+y} \left(\frac{k}{k_{eq}}\right)^2 \Phi = -\frac{3}{2y^2} w\Gamma, \end{aligned} \quad (5.175)$$

where k_{eq} is defined by (5.140). In the case $\Gamma = 0$, we have integrated this equation to obtain the solutions depicted in Fig. 5.9.

For non-vanishing entropy perturbation, one needs to know the actual behaviour of Γ in order to solve (5.140). As we shall show later, for large wavelengths, Γ remains essentially constant, so that the results of Fig. 5.9 are then a good approximation for these modes in the adiabatic case.

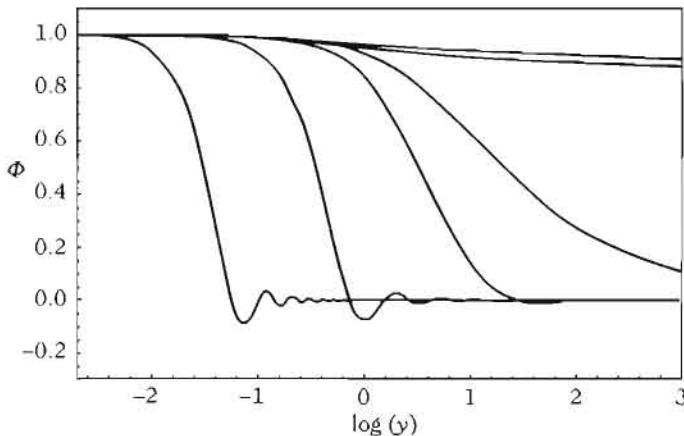


Fig. 5.9 Evolution of the gravitational potential in a Universe dominated by a mixture of matter and radiation, assuming that $\Gamma = 0$ for $k/k_{eq} = 10^{-3}, 0.1, 1, 2, 10, 100$. This figure should be compared with Fig. 5.12 where the time variation of the entropy is taken into account.

It is interesting to note that in the regime $k/k_{eq} \ll 1$, this equation can be solved analytically. Indeed, setting $\phi = y^3\Phi/\sqrt{1+y}$, we easily obtain $d\Phi/dy \propto y^2(4+3y)/(1+y)^{3/2}$ which can be integrated to give

$$\Phi = \frac{\Phi_0}{10y^3} \left(16\sqrt{1+y} + 9y^3 + 2y^2 - 8y - 16 \right). \quad (5.176)$$

For a super-Hubble mode, we recover $\Phi(y \gg 1)/\Phi_0 = 9/10$, which is a particular case of the result (5.159).

5.3.4 Mixture of several fluids

In the actual cosmological context, we should consider several fluids, for instance, baryons, dark matter, radiation, etc. These different fluids interact through gravity, but can also be coupled by non-gravitational forces. Figure 5.10 summarizes the main components that should be considered and their potential interactions.

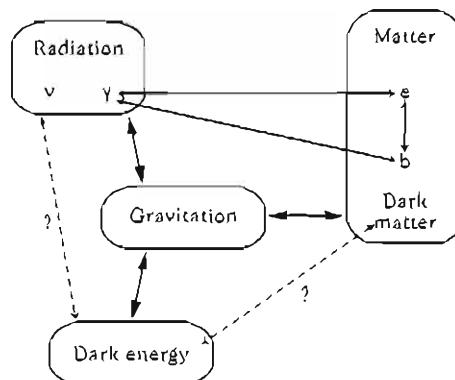


Fig. 5.10 The different cosmological fluids and their interactions. Electrons and baryons are coupled via the Coulomb force and photons interact with charged matter via Compton scattering. Dark matter only interacts gravitationally with ordinary matter. It is not excluded that dark energy interacts with ordinary matter or with dark matter in an as yet unknown way.

5.3.4.1 Description of the interactions

In the most general coupled case, the conservation of the energy-momentum tensor is only valid for the total energy-momentum, whereas the individual energy-momentum tensors satisfy conservation equations of the form

$$\nabla_\mu T_a^{\mu\nu} = Q_a^\nu, \quad (5.177)$$

where Q_a^ν represents the forces acting on the component a of the fluid. These terms must satisfy the constraint

$$\sum Q_a^\nu = 0, \quad (5.178)$$

which is simply the action-reaction law. It can be derived from the Bianchi identities that imply the conservation of the total energy-momentum tensor.

To lowest order, space-time symmetries imply that

$$Q_a^\mu = (-aQ_a, 0), \quad (5.179)$$

so that $Q_\mu^a Q_\nu^a \bar{g}^{\mu\nu} = -Q_a^2$. Note that $(\bar{g}_{\mu\nu} + \bar{u}_\mu \bar{u}_\nu) Q_a^\mu = 0$, so that the symmetries of the background space-time imply that there cannot exist any force between the fluids, at the background level. Taking the couplings into account, the conservation equations generalize to

$$\rho'_a + 3\mathcal{H}(1+w_a)\rho_a = aQ_a, \quad (5.180)$$

and we find

$$w'_a = - \left[3\mathcal{H}(1+w_a) - \frac{Q_a}{\rho_a} \right] (c_a^2 - w_a), \quad (5.181)$$

where w_a and c_a^2 are, respectively, the equation of state and the sound speed of the component a .

At linear order in perturbations, Q_a^μ can be decomposed as

$$Q_a^\mu = Q_a u_a^\mu + F_a^\mu, \quad \text{with} \quad F_a^\mu u_\mu^a = 0. \quad (5.182)$$

We define $af_a^\mu = (\delta_\nu^\mu + u^\mu u_\nu)F_a^\nu$, so that $f_a^0 = 0$ and f_a^i can, as usual, be decomposed into a scalar and a vector as $f_a^i = \bar{f}_a^i + D^i f_a$. Under a gauge transformation of the form (5.57), these quantities transform as

$$\delta Q_a \rightarrow \delta Q_a + Q'_a T, \quad (5.183)$$

and f_a^i is gauge invariant, since it vanishes at the background level. Thus, we can define the quantity

$$\delta Q_a^N = \delta Q_a + Q'_a(B - E'), \quad (5.184)$$

which is gauge invariant. With these notations, the continuity equation (5.125) and the Euler equation (5.126) for the component a generalize as

$$\left(\frac{\delta_a^N}{1+w_a} \right)' = -(\Delta V_a - 3\Psi') - 3\mathcal{H} \frac{w_a}{1+w_a} \Gamma_a + \frac{a}{\rho_a + P_a} [\delta Q_a^N + Q_a (\Phi - \delta_a^N)], \quad (5.185)$$

and

$$V'_a + \mathcal{H}V_a = -\Phi - \frac{c_a^2}{1+w_a} \delta_a^c + \frac{w_a}{1+w_a} \left[\Gamma_a + \frac{2}{3}(\Delta + 3K)\bar{\pi}_a \right] + \frac{a}{\rho_a + P_a} (f_a - Q_a c_a^2 V_a). \quad (5.186)$$

5.3.4.2 System of fluids interacting gravitationally

We consider the simplified case in which a set of fluids interact only through gravitation. In this case, $Q^\mu = 0$ for each fluid and the energy-momentum tensor is thus the sum of the energy-momentum tensors

$$T^{\mu\nu} = \sum_a T_a^{\mu\nu}.$$

Each energy-momentum tensor can be decomposed as in (5.69) so that the total density and pressure can be defined as

$$\rho = \sum_a \rho_a, \quad P = \sum_a P_a, \quad (\rho + P)v^i = \sum_a (\rho_a + P_a)v_a^i. \quad (5.187)$$

We have

$$\Omega w = \sum_a \Omega_a w_a, \quad (5.188)$$

and

$$\Omega c_s^2 = \sum_a \frac{1+w_a}{1+w} \Omega_a c_a^2. \quad (5.189)$$

For the sake of simplicity, let us consider a system of two fluids. It will be described by two density contrasts, δ_a and δ_b , and two velocity fields, v_a^i and v_b^i . It is convenient to transform these variables into a set of variables, δ and v , describing the properties of the total fluid and a set of variables S_{ab} and V_{ab}^i , describing their relative properties. The definitions (5.187) imply that

$$\Omega\delta = \sum_a \Omega_a \delta_a \quad \text{and} \quad \Omega(1+w)v = \sum_a \Omega_a (1+w_a)v_a, \quad (5.190)$$

for the scalar modes. These quantities satisfy the continuity equation (5.125) and the Euler equation (5.126). The entropy perturbation, Γ , is defined as before by the relation (5.78), so that, for such a mixture of fluids, it takes the form

$$\Omega w \Gamma = \sum_a w_a \Omega_a \Gamma_a + \sum_a (c_s^2 - c_a^2) \Omega_a \delta_a. \quad (5.191)$$

The first term corresponds to the entropy contribution from the components of the mixture and the second term, which does not vanish even if $\Gamma_a = 0$ for each component, represents the entropy of mixing. One can check that this second contribution is gauge invariant.

The Einstein equations take the general form obtained previously but for the quantities describing the total fluid. In particular, both (5.117) and (5.118) can be rewritten in the form

$$(\Delta + 3K)\Psi = \frac{3}{2}\mathcal{H}^2 \sum_a \Omega_a \delta_a^c, \quad (5.192)$$

$$\Psi - \Phi = 3\mathcal{H}^2 \sum_a \Omega_a w_a \bar{\pi}_a. \quad (5.193)$$

Now, S_{ab} and V_{ab}^i are defined, respectively, by

$$S_{ab} = \frac{\delta_a}{1+w_a} - \frac{\delta_b}{1+w_b} \quad \text{and} \quad V_{ab} = v_a - v_b, \quad (5.194)$$

which are, by construction, gauge invariant. Using the relations (5.190), we easily obtain the inverse relation

$$\left(\frac{\Omega_b}{1+w_a} + \frac{\Omega_a}{1+w_b} \right) \delta_a = \frac{\Omega}{1+w_b} \delta + \Omega_b S_{ab}, \quad (5.195)$$

and the analogous expression for δ_b is obtained by interchanging a and b .

The evolution equations for the quantities S_{ab} and V_{ab} are obtained by combining the continuity equation in the form (5.127) and the Euler equation (5.130) for each fluid. We get

$$S'_{ab} = -\Delta V_{ab} - 3\mathcal{H}\Gamma_{ab}, \quad (5.196)$$

$$\begin{aligned} V'_{ab} = & -\mathcal{H}V_{ab} - (c_a^2 - c_b^2) \frac{\delta^c}{1+w} - \left[c_a^2(1+w_b) \frac{\Omega_b}{\Omega} + c_b^2(1+w_a) \frac{\Omega_a}{\Omega} \right] \frac{S_{ab}}{1+w} \\ & - \left[\frac{2}{3} (\Delta + 3K) \bar{\pi}_{ab} + \Gamma_{ab} \right], \end{aligned} \quad (5.197)$$

with the definitions

$$\Gamma_{ab} \equiv \frac{w_a}{1+w_a} \Gamma_a - \frac{w_b}{1+w_b} \Gamma_b, \quad (5.198)$$

$$\bar{\pi}_{ab} \equiv \frac{w_a}{1+w_a} \bar{\pi}_a - \frac{w_b}{1+w_b} \bar{\pi}_b. \quad (5.199)$$

To completely solve the evolution of a system of two interacting fluids, we thus need to solve the evolution (5.143) for the gravitational potential, where Γ is now given by the relation (5.191), together with the pair of equations (5.196) and (5.197): the relation (5.192) then gives δ^c .

5.3.4.3 Adiabatic and isocurvature modes

Two different types of initial conditions can be distinguished for a system of two fluids: we can decide that the entropy perturbation vanishes initially (adiabatic initial conditions), or, on the contrary, arrange that the gravitational potential initially vanishes, i.e. ‘balance’ the density perturbations so that $\delta^c = 0$ (isocurvature initial conditions). We thus define the two limiting regimes

– *adiabatic*: it is assumed that $\Gamma = S_{ab} = 0$, from which we deduce that

$$\frac{\delta_a}{1+w_a} = \frac{\delta_b}{1+w_b}, \quad (5.200)$$

– *isocurvature*: in this case, we impose that $\Psi = 0$ and thus we have

$$\delta^c = 0, \quad \Omega_a \delta_a^c + \Omega_b \delta_b^c = 0. \quad (5.201)$$

Since the evolution equations are linear, any system can always be decomposed into a linear combination of an initially adiabatic system and an initially isocurvature system.

5.3.4.4 Mixture of two perfect fluids

The discussion in the previous section is very general. In practice, we are interested in simpler special cases. Let us consider that of a mixture of two perfect fluids ($\Gamma_a = 0$, $\pi_a = 0$) with purely gravitational interactions. Using (5.188) and (5.189), the relation (5.191) reduces to

$$w\Gamma = -\frac{\Omega_a \Omega_b (1+w_a)(1+w_b)}{\Omega^2(1+w)} (c_a^2 - c_b^2) S_{ab},$$

so that a closed system can be written for δ^c and S_{ab} , namely

$$\Phi'' + 3\mathcal{H}(1+c_s^2)\Phi' + [2\mathcal{H}' + (\mathcal{H}^2 - K)(1+3c_s^2)]\Phi - c_s^2 \Delta \Phi = \frac{3}{2}\mathcal{H}^2 \frac{\Omega_a \Omega_b (1+w_a)(1+w_b)}{\Omega(1+w)} (c_a^2 - c_b^2) S_{ab}, \quad (5.202)$$

$$(\Delta + 3K)\Phi = \frac{3}{2}\mathcal{H}^2 \Omega \delta^c, \quad (5.203)$$

$$S_{ab}'' + \mathcal{H} S_{ab}' - \frac{\Omega_a (1+w_a) c_b^2 + \Omega_b (1+w_b) c_a^2}{\Omega(1+w)} \Delta S_{ab} = (c_a^2 - c_b^2) \frac{\Delta \delta^c}{1+w}. \quad (5.204)$$

This system reduces to a system of two coupled second-order differential equations if the gravitational potential is eliminated from (5.202) by differentiating (5.203).

5.4 Power spectrum of density fluctuations

We will now be interested in the power spectrum of dark-matter density fluctuations. For this, and for the sake of simplicity, we assume that all matter is dark (i.e. $\Omega_{b0} \ll \Omega_{c0} \simeq \Omega_{m0} \simeq 1$). Moreover, we restrict ourselves to a flat Universe ($K = 0$ and $\Omega = 1$) with no cosmological constant. The Universe thus contains, by construction, only dark matter and radiation. We choose the normalization $a_0 = 1$.

As before, we introduce the scale factor y normalized at equality [see (5.136)] so that the Hubble law takes the form (5.139). The value of y today is

$$y_0 = \frac{a_0}{a_{eq}} = 1 + z_{eq} \simeq 2.4 \times 10^4 h^2$$

[see (4.8)]. The Friedmann equation evaluated today makes it possible to compute the numerical value of the mode entering the Hubble radius at the time of matter-radiation equality (5.140), namely

$$k_{eq} = H_0 \sqrt{2 \frac{\Omega_{m0} a_0}{a_{eq}}} = 0.072 \Omega_{m0} h^2 \text{Mpc}^{-1}, \quad (5.205)$$

corresponding to a wavelength

$$k_{eq}^{-1} = \mathcal{H}_{eq}^{-1} = \frac{14}{\Omega_{m0} h^2} \text{Mpc}. \quad (5.206)$$

In general, a given mode k enters the Hubble radius when $k = \mathcal{H}$, and the Friedmann equation (5.139) allows us to determine that this occurs at a redshift

$$y_* (\bar{k}) = \frac{1 + \sqrt{1 + 8\bar{k}}}{4\bar{k}^2}, \quad 1 + z_* (\bar{k}) = \frac{y_0}{y_* (\bar{k})}, \quad (5.207)$$

where we have introduced the reduced wave number

$$\bar{k} = \frac{k}{k_{\text{eq}}} \quad (5.208)$$

for further notational convenience.

5.4.1 Two equivalent approaches

We can follow two approaches to study this physical system. The first is to consider the 4 equations of evolution for the density contrasts and the velocity perturbations of each fluid, and the Poisson equation. The second is to use the results of the previous section for the total density contrast and the entropy perturbations. We will use one or the other formulation depending on the situation.

5.4.1.1 Evolution of the density contrasts

Our system is composed of a mixture of radiation and matter that only interacts gravitationally, so that the system of equations is deduced from (5.125) and (5.126). In Fourier modes, they reduce to

$$\delta_m^{N'} = k^2 V_m + 3\Phi', \quad (5.209)$$

$$\delta_r^{N'} = \frac{4}{3}k^2 V_r + 4\Phi', \quad (5.210)$$

$$V'_m = -\mathcal{H}V_m - \Phi, \quad (5.211)$$

$$V'_r = -\Phi - \frac{1}{4}\delta_r^N, \quad (5.212)$$

where we have used that $\Psi = \Phi$ since $\bar{\pi} = 0$. Another equation should be added to determine the gravitational potential and we choose to use the Poisson equation

$$-k^2\Phi = \frac{3}{2}\mathcal{H}^2 \left[\Omega_m \delta_m^N + \Omega_r \delta_r^N - 3\mathcal{H} \left(\Omega_m V_m + \frac{4}{3}\Omega_r V_r \right) \right], \quad (5.213)$$

or (5.119) in the form

$$\Phi' + \mathcal{H}\Phi = -\frac{3\mathcal{H}^2}{2} \left(\Omega_m V_m + \frac{4}{3}\Omega_r V_r \right). \quad (5.214)$$

These two equations can also be used in a combined way to express Φ in terms of δ_m^N and δ_r^N only.

5.4.1.2 Evolution of the gravitational potential and of the entropy

Using the expressions (5.174) to express c_s^2 and w , the system (5.202)–(5.204) reduces to

$$\ddot{\Phi} + \left(7 - \frac{1}{1+y} + \frac{8}{4+3y} \right) \frac{\dot{\Phi}}{2y} + \frac{\Phi}{y(1+y)(4+3y)} = \frac{2}{(4+3y)y^2} \left(\delta^c - \frac{yS}{1+y} \right), \quad (5.215)$$

$$\delta^c = -\frac{4}{3} \left(\frac{k}{k_{eq}} \right)^2 \frac{y^2}{1+y} \Phi, \quad (5.216)$$

$$\ddot{S} + \frac{3y+2}{2y(1+y)} \dot{S} = \frac{2}{4+3y} \left(\frac{k}{k_{eq}} \right)^2 \left(\delta^c - \frac{y}{1+y} S \right). \quad (5.217)$$

with $S = \delta_m - \frac{3}{4}\delta_r$. We have also assumed that for a Universe with Euclidean spatial sections and vanishing cosmological constant ($K = \Lambda = 0$), the matter and radiation density parameters can be rewritten simply in terms of the rescaled scale factor y as

$$\Omega_m = \frac{y}{1+y} \quad \text{and} \quad \Omega_r = \frac{1}{1+y}. \quad (5.218)$$

This system generalizes (5.175) to the case where the entropy perturbation is no longer neglected. The density contrast δ^N is then given by

$$\delta^N = \delta^c - 2(\Phi + y\dot{\Phi}), \quad (5.219)$$

and the density contrasts of each fluid are

$$\delta_m = \frac{3(1+y)\delta/4 + S}{1+3y/4}, \quad \text{and} \quad \delta_r = \frac{(1+y)\delta - yS}{1+3y/4}. \quad (5.220)$$

5.4.1.3 Initial conditions

To solve the systems (5.209)–(5.212) or (5.215)–(5.217), initial conditions for the perturbations need to be imposed (see, for example, Ref. [39]). These initial conditions are fixed deep within the radiation era ($y_i \ll 1$) and for super-Hubble modes ($k/\mathcal{H}_i \ll 1$). As seen earlier, we can distinguish two types of initial conditions:

- adiabatic: the super-Hubble modes are assumed constant (in other words the decaying mode has had enough time to decay) and, by definition, we have,

$$\Phi = \Phi_i, \quad \Phi' = 0, \quad S = 0, \quad S' = 0,$$

which fixes everything for the system (5.215)–(5.217). The Poisson equation implies that

$$\delta_i^c = -\frac{2}{3}x_i^2\Phi, \quad \delta_{r,i}^c = \delta_{m,i}^c = \frac{3}{4}\delta_i^c, \quad (5.221)$$

where $x_i \equiv k/\mathcal{H}_i$. (5.214) then implies that

$$kV_i = -\frac{1}{2}x_i\Phi, \quad V_{r,i} = V_{m,i} = V_i, \quad (5.222)$$

and (5.219) that

$$\delta_i^N = -2\Phi_i, \quad \delta_{r,i}^N = \delta_i^N, \quad \delta_{m,i}^N = \frac{3}{4}\delta_i^N, \quad (5.223)$$

which completely fixes the initial conditions for the system (5.209)–(5.212).

- *isocurvature*: we still assume the super-Hubble modes to be constant and the definition yields

$$\Phi = 0, \quad \Phi' = 0, \quad S = S_i, \quad S' = 0.$$

The Poisson equation implies that

$$\delta_i^C = 0, \quad \delta_{r,i}^C = -y_i S_i, \quad \delta_{m,i}^C = S_i, \quad (5.224)$$

so that (5.214) then implies

$$V_{r,i} = V_{m,i} = V_i = 0, \quad (5.225)$$

while (5.219) provides

$$\delta_i^N = \delta_i^C, \quad (5.226)$$

which again completely fixes the initial conditions for the system (5.215)–(5.217).

5.4.2 Different regimes

The behaviour of the solutions of these systems mainly depends on the value of the wave number k that determines the time at which the mode becomes sub-Hubble.

In order to have an idea of the orders of magnitudes, the following Table gives the values of y and of the redshift at the time a given mode becomes sub-Hubble for several selected values of k .

| $k (h^{-1} \text{ Mpc})$ | 10^3 | 10^2 | 10 | 1 | 0.1 | 10^{-2} | 10^{-3} |
|--------------------------|--------------------|--------------------|--------------------|----------------------|----------------------|--------------------|-------------------|
| y_* | 5×10^{-5} | 5×10^{-4} | 5×10^{-3} | 5.1×10^{-2} | 6.4×10^{-1} | 26 | 2.5×10^3 |
| z_* | 4.75×10^8 | 4.75×10^7 | 4.74×10^6 | 4.63×10^5 | 3.7×10^4 | 9.05×10^2 | 8.4 |

We therefore have four kinds of behaviour to take into account, depending on whether the mode is (i) super-Hubble today, (ii) has become sub-Hubble during the matter-dominated era, or (iii) during the radiation-dominated era, or (iv) whether it has always been sub-Hubble.

5.4.2.1 Super-Hubble modes

Let us first consider long-wavelength modes. In practice, one can neglect all the terms that involve a Laplacian, which are proportional to k^2 . Equation (5.196) then tells us that the entropy remains constant. Actually, we can also reach this conclusion by

noting that (5.216) implies that $\delta^c \propto k^2 \Phi$. In the set of equations (5.215)–(5.217), the terms $k^2 \Phi$ and $k^2 S$ are negligible in comparison to, respectively, $\dot{\Phi}$ and \ddot{S} . We then infer that in (5.217), $k^4 \Phi / k_{\text{eq}}^2 \ll \ddot{S}$, which implies that S is constant.

Note then that using the transformation (5.87) to introduce δ_r^c which is of the order of $\delta^c \propto k^2 \Phi$ in the radiation era, the two equations (5.211) and (5.212) imply that

$$V'_{rm} = -\mathcal{H}V_{rm} - \frac{1}{4}\delta_r^c,$$

and therefore that $V_{rm} \propto a^{-1}$ during the radiation era.

In the case of adiabatic initial conditions, $S = 0$ and the solution can be obtained analytically in the form (5.176). Deep enough into the radiation era, both the gravitational potential and the total density contrast are constant. In particular,

$$\delta^N(y) = -2\Phi(y), \quad \delta_r^N(y) = \frac{4}{3}\delta_m^N(y),$$

whatever y . In the radiation era, we obtain

$$\delta_r^N = -2\Phi(y \ll 1), \quad \delta_m^N(y) = -\frac{3}{2}\Phi(y \ll 1),$$

and in the matter era ($y \gg 1$)

$$\begin{aligned} \Phi(y \gg 1) &= \frac{9}{10}\Phi(y \ll 1), \\ \delta_r^N(y \gg 1) &= -\frac{12}{5}\Phi(y \ll 1), \quad \delta_m^N(y \gg 1) = -\frac{9}{5}\Phi(y \ll 1). \end{aligned} \quad (5.227)$$

To illustrate the effect of the initial conditions, let us now consider some isocurvature initial conditions and let us assume that initially $S = S_i \neq 0$ and $\Phi = 0$. As we have seen S remains constant. Deep in the radiation era, ($y \ll 1$), $\delta_r^N \sim \delta^N = 2\Phi = 0$ and thus $\delta_m^N = S_i$. Note then that for super-Hubble modes, since Φ and S are constant, $\Phi = -2S$ is a particular solution of (5.215). The general solution satisfies the initial condition $\Phi_i = 0$ and is thus obtained by adding this particular solution to the general solution (5.176) where the constant Φ_0 should be taken equal to $\Phi_0 = 2S_i$,

$$\Phi = \frac{2S_i}{10y^3} \left(16\sqrt{1+y} + 9y^3 + 2y^2 - 8y - 16 \right) - 2S_i. \quad (5.228)$$

In the matter-dominated era ($y \gg 1$), $\delta_m^N \sim \delta^N = -2\Phi$ and $\Phi(y \gg 1) = -S_i/5$. We conclude that, for $y \gg 1$

$$\Phi = -\frac{1}{5}S_i, \quad \delta^N \sim \delta_m^N = \frac{2}{5}S_i, \quad \delta_r^N = -\frac{4}{5}S_i. \quad (5.229)$$

5.4.2.2 Modes entering the Hubble radius in the matter-dominated era

As can be seen in Fig. 5.11, even though the mode is sub-Hubble from a redshift of the order of $z_* \sim 8.5$, it is always constant. Actually, to establish this result, we had

to assume that the Laplacian term was negligible in the evolution equation for the gravitational potential. This is legitimate as long as $c_s^2 k^2 \ll \mathcal{H}^2$, i.e. if

$$\frac{2y^2}{3(1+y)(1+3y/4)} \bar{k}^2 \ll 1. \quad (5.230)$$

For the modes represented in Fig. 5.11, $\bar{k} = 10^{-3}$ and this ratio is still equal to 1.7×10^{-4} today. One can check that this ratio is of order unity for $k \sim k_{\text{eq}}$ so that for all the modes entering the Hubble radius in the matter-dominated era, the

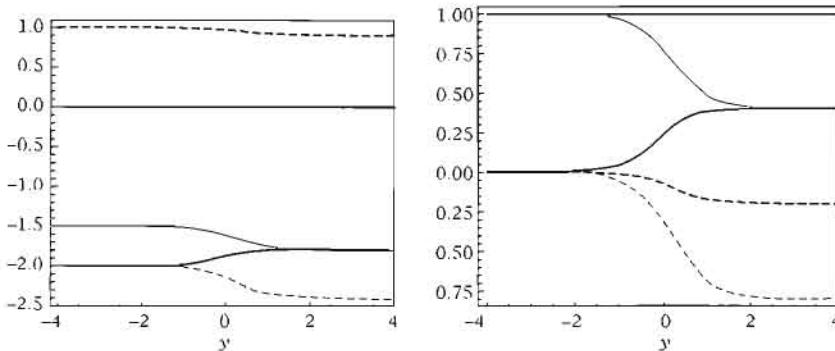


Fig. 5.11 Evolution of the total density contrast δ^N (thick solid line), the entropy S (thick dotted line) and gravitational potential (thick dashed line) as well as the two contrasts δ_r^N (thin dashed line) and δ_m^N (thin solid line) for a mode $\bar{k} = 10^{-3}$ assuming either adiabatic initial conditions ($\Phi_i = 1, S_i = 0$) [left] or isocurvature initial conditions ($\Phi_i = 0, S_i = -1$) [right].

gravitational potential remains constant, put aside the small variation due to the jump of the equation of state from $\frac{1}{3}$ to 0 (see Fig. 5.12). The Poisson equation then tells us that

$$\delta_m^c \propto a. \quad (5.231)$$

5.4.2.3 Modes entering the Hubble radius in the radiation-dominated era

When the Universe is dominated by radiation, the gravitational potential is determined by the density fluctuations of radiation. It is therefore given by the general solution (5.161) and its evolution is represented in Fig. 5.5. Thus, as soon as the modes becomes sub-Hubble during the radiation-dominated era, the gravitational potential is subject to damped oscillations and tends rapidly to 0.

The dark-matter fluid density perturbations are then obtained by solving their evolution equations with the gravitational potential determined by the solution (5.161),

$$\delta_m^{N''} + \mathcal{H}\delta_m^{N'} = 3\Phi'' + 3\mathcal{H}\Phi' - k^2\Phi, \quad (5.232)$$

where the Hubble parameter is given by $\mathcal{H} = \eta^{-1}$ deep enough into the radiation era. The solution of this equation is of the form

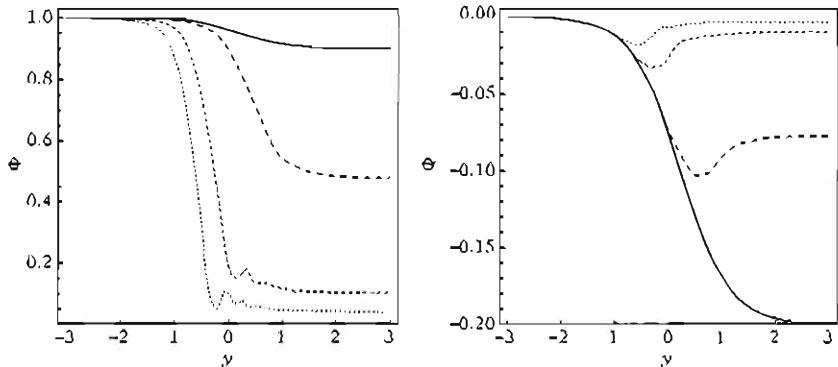


Fig. 5.12 Evolution of the gravitational potential for $k = 10^{-2}, 0.1, 0.5, 1 \text{ Mpc}^{-1}$ for adiabatic (left) and isocurvature (right) initial conditions. These figures should be compared with Fig. 5.9 for which the entropy was neglected.

$$\delta_m^N = A + B \ln(k\eta) + \delta_{\text{part}}^N, \quad (5.233)$$

where the particular solution is obtained in the integral form

$$\delta_{\text{part}}^N(k, \eta) = \int_0^\eta d\eta' [3\Phi''(k, \eta') + 3H\Phi'(k, \eta') - k^2\Phi(k, \eta')] \eta' \ln\left(\frac{\eta'}{\eta}\right). \quad (5.234)$$

When $y \ll 1$, $\delta^N \sim \delta_r^N \sim 2\Phi_i$. Assuming adiabatic initial conditions ($\delta_m^N = 3\delta_r^N/4$) then $\delta_m^N \sim 3\Phi_i/2$. It turns out that the contribution of the particular solution is negligible so that we are left with $A = 3\Phi_i/2$ and $B = 0$.

As shown by the solution (5.161), the gravitational potential decays rapidly as soon as the mode becomes sub-Hubble and is constant before. The integrand varies significantly only around $\eta' = 1/k$, so that the upper bound of the integral can be extended to infinity in the particular solution. This means that the above integral can be performed to give

$$\delta_{\text{part}}^N(k, \eta) = A + B \ln(k\eta), \quad (5.235)$$

where the constants

$$A = \int_0^\infty f(k\eta') k\eta' \ln(k\eta') dk\eta' \quad \text{and} \quad B = - \int_0^\infty f(k\eta') k\eta' dk\eta' \quad (5.236)$$

depend on the function f , itself given explicitly by

$$f(u) = 9\Phi_i u^{-5} \left[u(2u^2 - 27) \cos \frac{u}{\sqrt{3}} + \sqrt{3}(27 - 5u^2) \sin \frac{u}{\sqrt{3}} \right],$$

obtained by using the solution (5.161). Numerically, we obtain $A \simeq -6$ and $B \simeq 9$ so that the general solution takes the form

$$\delta_m^N(k, \eta) \simeq \Phi_i [-4.5 + 9 \ln(k\eta)]. \quad (5.237)$$

Obtaining the exact value of these coefficients would of course require a more careful calculation. To conclude, let us recall that

$$\delta_m^N \sim \delta_m^C \propto \ln a \quad (k\eta \gg 1); \quad \delta_m^N \propto a^0 \quad (k\eta \ll 1) \quad (5.238)$$

during the radiation era (see Fig. 5.13 for a numerical integration).

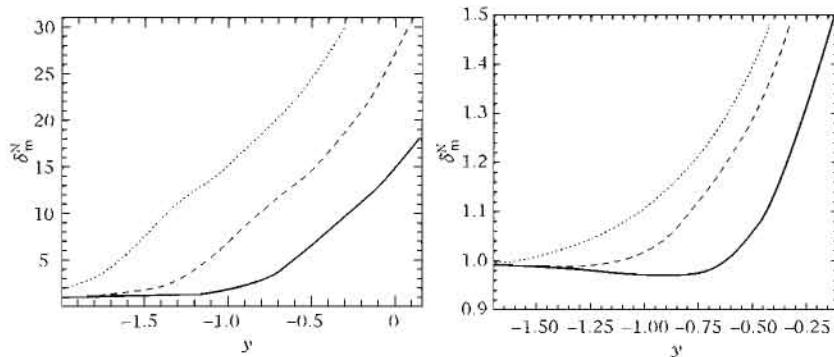


Fig. 5.13 Evolution of the matter density contrast δ_m^N for $k = 1 h^{-1} \text{ Mpc}^{-1}$ (solid line), $k = 5 h^{-1} \text{ Mpc}^{-1}$ (dashed line) and $k = 10 h^{-1} \text{ Mpc}^{-1}$ for adiabatic (left) and isocurvature (right) initial conditions. These modes become sub-Hubble during the radiation-dominated era.

5.4.2.4 Sub-Hubble modes

Let us now consider the modes that were sub-Hubble during the radiation era and let us try to evaluate how the perturbations of the matter fluid are affected by the radiation-matter transition. For these modes, we have seen (Fig. 5.6) that the density contrasts were the same independent of the gauge.

For any sub-Hubble mode during the radiation era, we have determined that δ_r is roughly constant, whereas from the previous section we know that $\delta_m \sim \ln a$. Even though ρ_m is small compared to ρ_r , the contribution of the matter density fluctuations can dominate over those of radiation in the Poisson equation, since

$$\frac{\rho_m \delta_m}{\rho_r \delta_r} \sim y \ln y.$$

We will thus focus on the regime where $\delta_r \ll \delta_m$. The total density contrast is then given by

$$\delta = \frac{y}{1+y} \delta_m,$$

and the Poisson equation takes the form

$$k^2 \Phi = -\frac{3}{4y} k_{\text{eq}}^2 \delta_m, \quad (5.239)$$

as long as the velocity contributions can be neglected. We now use the two equations (5.209)–(5.210) rewritten in terms of y . After differentiating (5.209) and eliminating

V_m using the equation (5.211) and \dot{V}_m again with (5.209), we obtain a second-order equation for δ involving Φ and its derivatives, which can then be eliminated with (5.239).

To simplify this equation, notice that we are interested in the short-wavelength modes, and thus $k/H \gg 1$, which implies that $k_{eq}^2 \ll k^2 y$. Neglecting all these terms, we obtain the equation, known under the name of the Mészáros equation [40],

$$\ddot{\delta}_m + \frac{2+3y}{2y(y+1)} \dot{\delta}_m - \frac{3}{2y(y+1)} \delta_m = 0. \quad (5.240)$$

This equation has an affine solution and the second solution can be obtained using the method of variation of the parameter. One can check that two independent solutions are

$$D_+(y) = y + \frac{2}{3} \quad \text{and} \quad D_-(y) = D_+(y) \ln \left(\frac{\sqrt{1+y} + 1}{\sqrt{1+y} - 1} \right) - 2\sqrt{1+y}. \quad (5.241)$$

In the matter-dominated era ($y \gg 1$), they correspond, respectively, to a growing and a decaying mode since

$$D_+(y) \propto y \quad \text{and} \quad D_-(y) \propto y^{-3/2}.$$

This solution can be matched to the solution (5.233)–(5.235) at the time where the mode enters the Hubble radius at $y_*(k)$ (see Fig. 5.14 for a numerical integration).

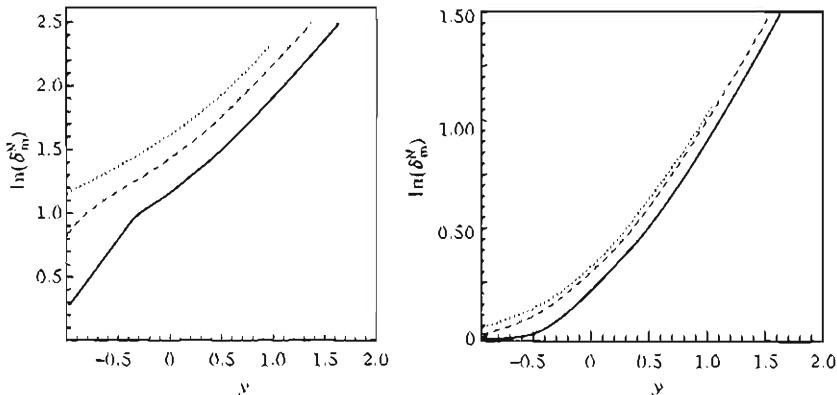


Fig. 5.14 Evolution of the matter-density contrast $\hat{\delta}_m^N$ for $k = 1 h^{-1} \text{ Mpc}^{-1}$ (solid line), $k = 5 h^{-1} \text{ Mpc}^{-1}$ (dashed) and $k = 10 h^{-1} \text{ Mpc}^{-1}$ for adiabatic (left) and isocurvature (right) initial conditions.

5.4.2.5 Summary

In the simplified example of a Universe containing only dark matter and radiation interacting gravitationally, the previous study shows that the behaviour of Φ and δ_m

depends on the value of the mode k and on the era during which the mode becomes sub-Hubble.

However, this analysis neglects four important effects:

1. Matter contains baryons that are coupled to radiation by Compton scattering. We describe this coupling and its effects on the perturbations in Chapter 6.
2. We have described radiation using a fluid approximation. This description is very rough and should actually be extended using a kinetic approach and radiation should be described by a Boltzmann equation. Such a description implies a hierarchy of equations, of which the fluid equations are the two first moments. This will be described in detail in Chapter 6 that is dedicated to the study of the cosmic microwave background.
3. The recent evolution of perturbations can be affected by the existence of curvature or of a cosmological constant. These effects were illustrated in Figs. 5.7 and 5.8.
4. Due to the presence of neutrinos, $\Phi \neq \Psi$, which changes, for instance, the ratio 9/10 for the variation of the gravitational potential on large scales.

5.4.2.6 Transfer function

This simplified study, nevertheless, allows us to understand the general shape of the large-scale structure power spectrum. To see this, define the transfer function

$$T(k, a) = \frac{\delta(k, a) D_+(a_i)}{\delta(k, a = a_i) D_+(a)}, \quad (5.242)$$

where a_i is an initial time and $D_+(a)$ the growth function (5.24). Note that a_i is arbitrary as long as it corresponds to a time before any scale of interest has become sub-Hubble. This transfer function characterizes the modification of the initial matter power spectrum during the evolution and depends on the cosmological parameters and on the matter content of the Universe. The term D_+ has been added by hand because during the matter era

$$\Delta\Phi = \frac{3}{2} H_0^2 \Omega_{m0} \frac{D_+(a)}{a} \frac{\delta}{D_+(a)}$$

and the factor $D_+(a)/a$ is approximately equal to unity during the matter era (Fig. 5.1). The definition (5.242) implies that

$$\Phi(k, a) = T(k, a) \Phi(k, a_i).$$

The observable modes today were initially super-Hubble and the theoretical models of the primordial Universe allow us to predict the matter power spectrum of these modes. This implies that the gauge in which the initial density contrast $\delta(k, a = a_i)$ is considered should be clearly specified. Actually, since for these modes Φ is constant, we will have the same transfer function if we use δ^c in the definition (5.242) since the two density contrasts are related to the gravitational potential by the same Poisson equation and we thus have

$$P_\Phi(k, a_0) = P_\Phi(k, a_i) T^2(k, a_0), \quad (5.243)$$

and

$$P_\delta(k, a_0) = P_\delta(k, a_i) T^2(k, a_0) \left[\frac{D_+(a_0)}{D_+(a_i)} \right]^2, \quad (5.244)$$

since $P_\delta(k, a_i) = k^4 P_\Phi(k, a_i)$.

The approximate shape of the transfer function can be deduced from our previous analysis (see Fig. 5.15). Indeed, neglecting the logarithmic growth during the radiation era, the modes that became sub-Hubble during this era are frozen and have not grown between the time they entered the Hubble radius and equality. Their relative amplitude compared to the long-wavelength modes (that is $k < k_{\text{eq}}$) is thus lowered by a factor $(k_{\text{eq}}/k)^2$. We thus expect the limiting behaviour to be

$$T(k, a_0) \sim 1 \quad (k \ll k_{\text{eq}}), \quad T(k, a_0) \sim \left(\frac{k_{\text{eq}}}{k} \right)^2 \quad (k \gg k_{\text{eq}}). \quad (5.245)$$

This result should be compared with Fig. 5.16 obtained from a numerical computation that includes in particular baryons and their coupling to the radiation. Note also that the transfer function will depend on the kind of initial conditions in the primordial Universe (adiabatic or isocurvature). Different approximations for this function exist in the literature [19]. As an example, for a cold dark matter model, one has

$$T(q) = \frac{\ln(1 + 2.34q)}{2.34q} \left[1 + 3.89q + (16.1q)^2 + (5.46q)^3 + (6.71q)^4 \right]^{-1/4}, \quad (5.246)$$

$$T(q) = (5.6q)^2 \left\{ 1 + \left[15q + (0.9q)^{3/2} + (5.6q)^2 \right]^{1.24} \right\}^{-1/1.24}, \quad (5.247)$$

with $q = k/(\Gamma \text{Mpc}^{-1})$ and $\Gamma = \Omega_{m0} h^2$, respectively, for adiabatic and isocurvature initial conditions.

5.4.3 Some refinements

To close this chapter, we give the complete set of equations, in the fluid approximation, for a Universe containing photons, neutrinos, dark matter and baryons. We encourage the reader to write a program for this system and to try to reproduce the curves of this chapter and to understand the different effects neglected in the previous presentation.

5.4.3.1 Dark matter

Dark matter is governed by the equations used all through this chapter,

$$\delta_c^{N'} = k^2 V_c + 3\Psi', \quad (5.248)$$

$$V_c' = -\mathcal{H}V_c - \Phi, \quad (5.249)$$

where we have maintained the distinction between Φ and Ψ .

5.4.3.2 Baryons

Electrons and nuclei have the same density contrast since the electromagnetic forces requires the charge density to be extremely close to zero. Their conservation equations

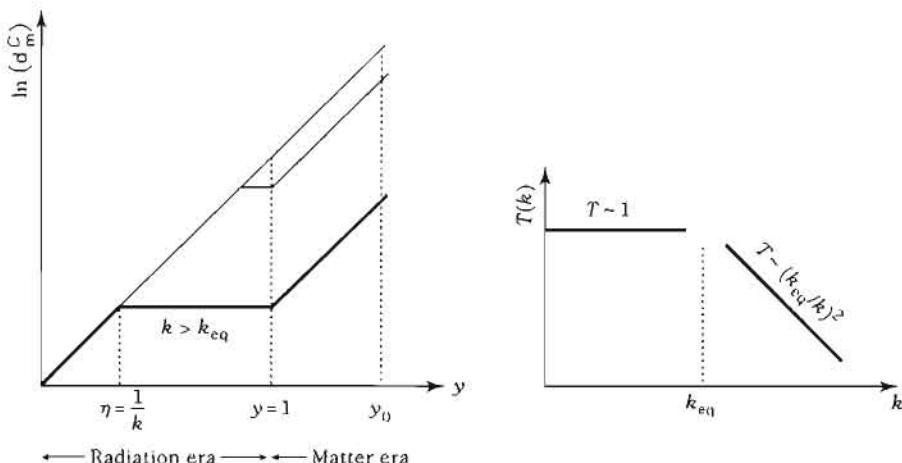


Fig. 5.15 (left): The modes that become sub-Hubble during the radiation era ($k > k_{eq}$) remains almost constant (neglecting the logarithmic growth) from $\eta \sim 1/k$ to $\eta \sim 1/k_{eq}$. (right): The amplitude of the density contrast on these scales suffers from a lack of growth of the order of $(k_{eq}/k)^2$. For $k > k_{eq}$ we thus expect the transfer function to behave as $T \propto (k_{eq}/k)^2$, while it remains of order unity for $k < k_{eq}$.

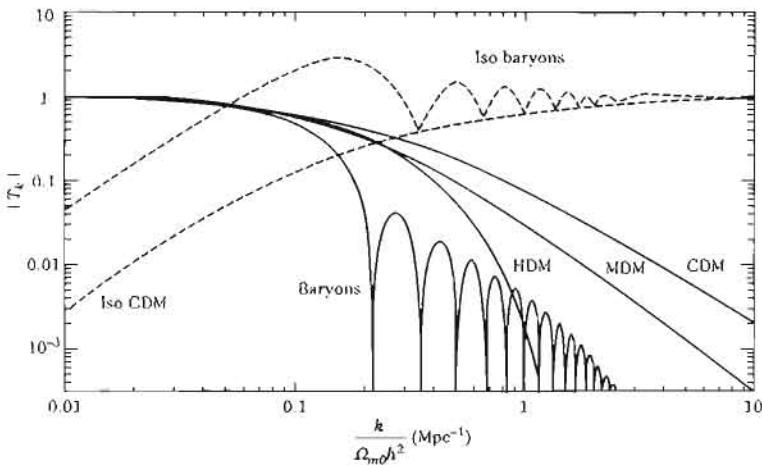


Fig. 5.16 Transfer function for different scenarios. The general form of this transfer function can be understood from the analytical study we have carried out, cf. (5.245). From Ref. [8].

are the same as for dark matter but their Euler equation will be different because the Thomson scattering couples them to photons. The momentum exchange between the baryon and the photon fluids should thus be taken into account. It can be shown [41] that the force that enters the Euler equation is

$$\frac{1}{\rho_b + P_b} f_b = \frac{4}{3} \frac{\rho_\gamma}{\rho_b} n_e \sigma_T (V_\gamma - V_b), \quad (5.250)$$

where σ_T is the Thomson scattering cross-section and n_e the free-electron density.

The speed of sound does not vanish and can be computed assuming that we have a monoatomic gas of mean molecular mass μ (it is actually the mean mass of electrons, protons and helium atoms) at a temperature T_b

$$c_b^2 = \frac{k_B T_b}{\mu} \left(1 - \frac{1}{3} \frac{d \ln T_b}{d \ln a} \right), \quad (5.251)$$

neglecting the time variation of μ (actually $\dot{\mu}$ is important only during the recombination period for which the baryons contribute very little to the pressure of the photon-baryon fluid, so this turns out to be a good approximation). To obtain the temperature evolution equation, let us start from the first law of thermodynamics (see Ref. [42]) $\delta Q = \frac{3}{2} \delta(P_b/\rho_b) + P_b \delta(1/\rho_b)$ with the heat variation rate given by $Q' = 4(\rho_\gamma/\rho_b) a n_e \sigma_T (T_\gamma - T_b)$. We conclude that the evolution of the baryon temperature is governed by

$$T'_b = -\mathcal{H} T_b + \frac{8}{3} \frac{\mu}{m_e} \frac{\rho_\gamma}{\rho_b} a n_e \sigma_T (T_\gamma - T_b). \quad (5.252)$$

Since $P_b \ll \rho_b$ one can neglect w_b and c_b^2 in the evolution equations apart from the term $c_b^2 \Delta \delta_b^N$. We eventually obtain

$$\delta_b^{N'} = k^2 V_b + 3\Psi', \quad (5.253)$$

$$V'_b = -\mathcal{H} V_b - \Phi - c_b^2 \delta_b^N + \frac{4}{3} \frac{\rho_\gamma}{\rho_b} a n_e \sigma_T (T_\gamma - T_b). \quad (5.254)$$

5.4.3.3 Photons

As will be seen in the next chapter, radiation should be described using a Boltzmann equation, which will have the effect of replacing the photon and neutrino propagation equations by two hierarchies of equations. Here, we only refine our fluid description by adding an anisotropic pressure term, π_γ , and the coupling to the baryon fluid

$$\delta_\gamma^{N'} = \frac{4}{3} k^2 V_\gamma + 4\Psi', \quad (5.255)$$

$$V'_\gamma = -\frac{1}{4} \delta_\gamma^N - \Phi + k^2 \sigma_\gamma + a n_e \sigma_T (T_\gamma - T_b). \quad (5.256)$$

We now have a force term stemming from the effect of the baryon fluid and $\sigma_\gamma \equiv \pi_\gamma/6$. We must accept without proof at this stage that the evolution equation for σ_γ is

$$\sigma'_\gamma = -\frac{4}{15} V_\gamma - \frac{9}{5} a n_e \sigma_T \sigma_\gamma. \quad (5.257)$$

Actually, (5.256) contains the contribution from other multipoles of the Boltzmann hierarchy and the fluid version is thus only a truncation of this hierarchy.

5.4.3.4 Neutrinos

As stated above, neutrinos should also be described by a Boltzmann equation. They follow equations identical to those describing photons apart from the absence of coupling with the baryons. We thus have

$$\delta_\nu^N' = \frac{4}{3} k^2 V_\nu + 4\Psi', \quad (5.258)$$

$$V_\nu' = -\frac{1}{4} \delta_\nu^N - \Phi + k^2 \sigma_\nu. \quad (5.259)$$

with $\sigma_\nu \equiv \pi_\nu/6$. As for the photon, we accept without proof that the evolution equation of σ_ν is

$$\sigma_\nu' = -\frac{4}{15} V_\nu. \quad (5.260)$$

The more general case of massive neutrinos is described in detail in Ref. [44].

5.4.3.5 Einstein equations

The Einstein equations reduce to

$$\Psi - \Phi = 6 \frac{\mathcal{H}^2}{k^2} (\Omega_\gamma \sigma_\gamma + \Omega_\nu \sigma_\nu) \quad (5.261)$$

and

$$-k^2 \Psi = \frac{3}{2} \mathcal{H}^2 \left[\sum_i \Omega_i \delta_i^N - \mathcal{H} \sum_i (1 + w_i) \Omega_i V_i \right]. \quad (5.262)$$

5.4.3.6 Initial conditions

Just as in the previous sections, we must fix the initial conditions for the super-Hubble modes in the radiation era. For adiabatic initial conditions, we obtain (see, for instance, Ref. [39])

$$\delta_\gamma^N(\eta_i) = \delta_\nu^N(\eta_i) = \frac{4}{3} \delta_b^N(\eta_i) = \frac{4}{3} \delta_c^N(\eta_i) \quad \text{and} \quad V_\gamma(\eta_i) = V_\nu(\eta_i) = V_b(\eta_i) = V_c(\eta_i). \quad (5.263)$$

As before, we have

$$\delta_\gamma^N(\eta_i) = -2\Phi(\eta_i), \quad kV_\gamma(\eta_i) = -\frac{1}{2} k\eta_i \Phi(\eta_i), \quad (5.264)$$

from which we deduce that

$$\sigma_\nu(\eta_i) = \frac{1}{15} (k\eta_i)^2 \Phi, \quad \sigma_\gamma(\eta_i) = 0, \quad (5.265)$$

and

$$\Psi = \left(1 + \frac{2}{5} R_\nu \right) \Phi, \quad (5.266)$$

with $R_\nu = \rho_\nu / (\rho_\nu + \rho_\gamma)$.

5.4.3.7 Tight-coupling approximation

Note that before decoupling, $V_b = V_\gamma$ as the Compton scattering strongly couples both species. After decoupling, the Universe becomes almost neutral and both fluids are decoupled. This system can thus be solved in the ‘tight-coupling’ limit where one goes from one regime to the other almost instantaneously (see, for instance, Ref. [43]). Before decoupling we thus have

$$V_b = V_\gamma = V, \quad V' = -\frac{3\Omega_b}{4\Omega_\gamma + 3\Omega_b} \mathcal{H}V - \frac{\Omega_\gamma}{4\Omega_\gamma + 3\Omega_b} \delta_\gamma^N - \Phi. \quad (5.267)$$

5.4.3.8 Conclusions

The system of equations given in detail in this section leads to a more realistic picture of the evolution of cosmological perturbations. While remaining in the framework of a fluid description, it now takes into account the departures from a perfect fluid for photons and neutrinos via the introduction of their anisotropic pressure. The two gravitational potentials are thus no longer equal. The coupling between baryons and photons has been included, but its derivation requires a kinetic approach.

This kinetic description of radiation will be discussed in the next chapter in which we will not only justify the unproved equations but also provide a more precise version of these equations. One should also describe the dynamics of recombination to have access to the time variation of the electronic density and go beyond the strong-coupling hypothesis.

The resolution of such a system allows us to obtain the transfer function of Fig. 5.16 to a good precision for a large class of cosmological models.

5.5 The large-scale structure of the Universe

Up to now this chapter was essentially focused on the linear regime. In Section 5.1.4 we have alluded to the slightly non-linear regime. We now briefly describe what observations tell us on the matter power spectrum.

5.5.1 Observing the large-scale structure

5.5.1.1 Galaxy catalogues

The measure of the matter power spectrum is mainly based on the construction of large galaxy catalogues. Recently two large catalogues, SDSS (Sloan Digital Sky Survey) [45] and 2dFGRS (2-degree field Galaxy Redshift Survey) [46], have drastically increased both the number of observed objects and the sky coverage. SDSS covers around a quarter of the sky and measures the position and absolute brightness of around 100 million celestial objects and the distance of more than one million galaxies; 2dFGRS determines the position and redshift of more than 250 thousand galaxies. These catalogues probe the structure of the Universe up to a redshift of around 0.3. A sample of the catalogue constructed by 2dFGRS is presented in Fig. 5.17.

The link between these observations and the three-dimensional power spectrum of dark matter is not direct. Indeed, only luminous matter, i.e. baryonic, can be observed. In order to derive any relevant conclusion as to the total matter distribution, we

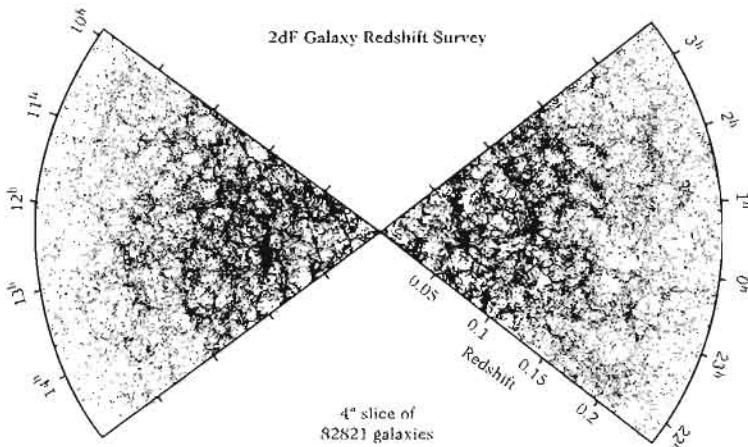


Fig. 5.17 The distribution of galaxies in a 4° wide range of the catalogue 2dfGRS. These observations of the large-scale structure of the Universe are at the basis of the measurements of the matter power spectrum.

should convince ourselves that this luminous matter is a good tracker of the dark-matter distribution. This implies that the density contrast of luminous matter should be proportional to that of dark matter. This coefficient of proportionality, b , is called the bias parameter [12, 18],

$$\delta_{\text{luminous}} = b\delta_c. \quad (5.268)$$

The validity of this hypothesis, the possibility that b may depend on scale and numerous other problems have been studied numerically and observationally [12]. Another problem lies in the fact that observations determine with a great precision the position of objects in the sky and their redshift. In redshift space, the radial separation includes information not only on the position of the galaxy but also on its velocity. Thus, at small scales, a spherical structure in gravitational collapse would appear, in the redshift space, as an ellipse flattened in the radial direction. This distortion in redshift space due to proper motion, must be corrected.

5.5.1.2 Constraints on the power spectrum

The large galaxy catalogues allow for the reconstruction of the matter power spectrum over around two orders of magnitude. The linear spectrum can be obtained from the anisotropies of the cosmic microwave background (Chapter 6).

On smaller scales, new techniques, such as weak gravitational lensing (Chapter 7) and the study of Lyman- α forests that allow for the reconstruction of the distribution of extragalactic gas, enable us to extend the measure by almost two orders of magnitudes in wavelength. Note that the use of gravitational lensing has the advantage of avoiding the bias problem. Indeed, this method directly measures the distribution of the gravitational potential (see Chapter 7 for details).

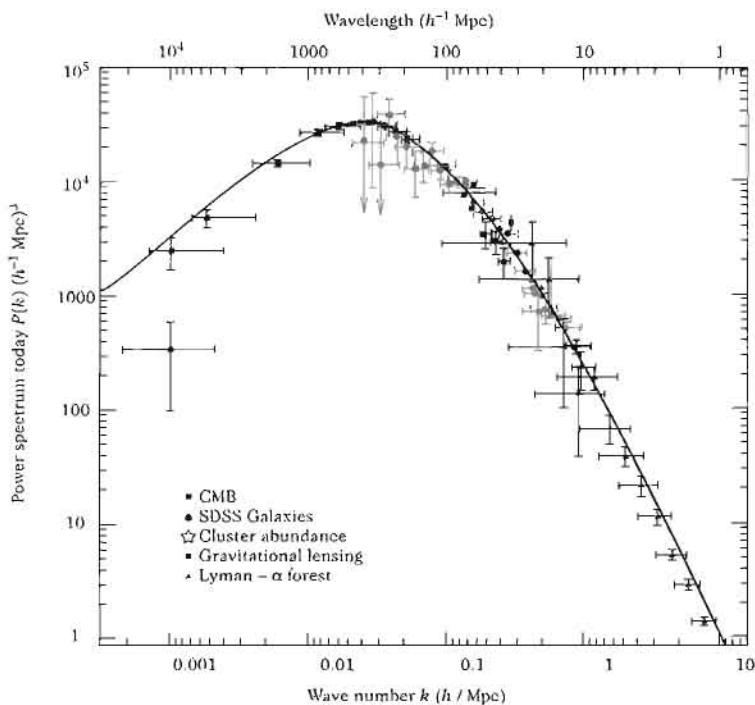


Fig. 5.18 The matter power spectrum compared with various observations from linear scales (cosmic microwave background) to non-linear scales (galaxy catalogues, here SDSS), gravitational lensing and Lyman- α forests. From Ref. [47].

To conclude, the matter power spectrum can be reconstructed over around five orders of magnitude. Figure 5.18 summarizes the state of observations and the reconstruction of the power spectrum performed using these different methods. Note however, that, although impressive at first sight, this figure cannot be directly deduced from observations. Indeed, for small scales ($k \gtrsim 0.1 h/\text{Mpc}$), only the power spectrum in the non-linear regime is observed. The linear reconstruction thus depends on a ‘mapping’ between the linear and non-linear regimes (see Fig. 5.19). Moreover, the reconstruction of the power spectrum from the Lyman- α forests is not direct either and, until further data is gathered, should be taken with caution.

5.5.1.3 Baryon acoustic oscillations

As we have already explained, during the radiation era photons and baryons are coupled. This implies that the photon–baryon plasma undergoes acoustic oscillations. The radiation pressure competes with gravitation and sets up oscillations in the photon fluid (see Fig. 5.5). The tight coupling between electrons, baryons and photons due to the Compton scattering then causes the baryons to oscillate in phase with the radiation. Thus, the whole plasma oscillates due to these sound waves.

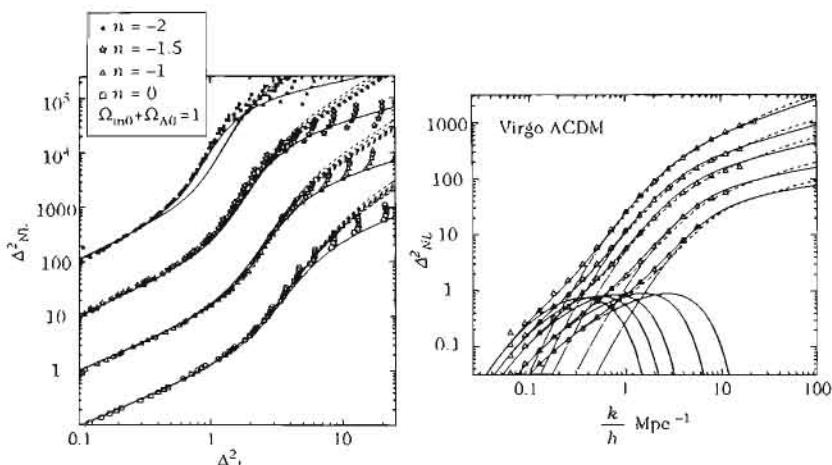


Fig. 5.19 (left): Comparison between the analytic mapping proposed by Ref. [51] and numerical simulations [53]. For each model, five epochs are represented. From bottom to top, $1/(1+z) = 0.6, 0.7, 0.8, 0.9, 1$. (right): The power spectrum in the non-linear regime at five epochs, $z = 0, 0.5, 1.0, 2.0, 3.0$ from bottom to top. The solid line corresponds to the mapping of Ref. [53] and the dotted line to that of Ref. [52]. From Ref. [53].

As we shall see in the next chapter, the imprint of these sound waves on the cosmic microwave background has been detected. These waves are related to a characteristic scale that corresponds to the sound horizon.

These oscillations are also imprinted in the power spectrum of galaxies. Moreover, their amplitude is suppressed by a factor of order $\Omega_b \sim 0.05$ since they are imprinted only in the baryon fluid and not in the dark matter. It was recently detected [48] in the correlation function of luminous red galaxies from the SDSS survey. This correlation exhibits a peak at a scale of order $100 h^{-1} \text{ Mpc}$ at a redshift of about $z \sim 0.35$.

5.5.2 Structures in the non-linear regime

The matter equations of evolution are non-linear and we have only solved them in the linear regime. To end this chapter, we present some approaches to the study of this regime that completes the analysis of Section 5.1.4.

5.5.2.1 Hierarchical model of structure formation

To understand the non-linear regime, it is crucial to determine when a scale enters this regime. If the density contrast was completely frozen during the radiation era, then all modes would become non-linear at the same time since the growth of perturbations in the matter era is independent of the wavelength. However, there is a small growth of perturbations during the radiation era (Figs. 5.13 and 5.14). The earlier a wavelength becomes sub-Hubble, the more it is amplified, so that small scales will be the first ones to become non-linear. The smallest structures are therefore the oldest. This is what is called a *hierarchic mechanism*.

At matter-radiation equality, the amplitude of perturbations is typically of the order of 10^{-5} . The small wavelengths enter the non-linear regime when $\delta_c(k, t) \sim 1$ and start forming structures. If, at this epoch, the Universe had been observed with a resolution larger than these scales, it would seem very homogeneous, so the density field can always be described by a continuous field in the linear regime.

The process continues and larger and larger wavelengths enter the non-linear regime. So, the small overdense regions (protogalaxies) form first before merging to form larger and larger galaxies and then clusters. The scale of non-linearity today is of the order of 20 to 40 Mpc, which is only one order of magnitude larger than the size of galaxy clusters (around 2 Mpc). So clusters are not distributed in a homogeneous way and form large structures, such as filaments, separated by voids the characteristic size of which is of the order of the non-linearity scale. At this scale, the structure of the Universe has the morphology of a sponge, but smoothed over a larger scale, it is homogeneous again.

5.5.2.2 Spherical collapse model

A simple analytical model of the non-linear evolution of a structure can be constructed by considering a spherical overdense region. To study its global evolution, it is not necessary to know the exact density profile since thanks to the Gauss theorem, only the mean density contained in a sphere of given radius is necessary to determine the evolution of this sphere.

In a matter-dominated Universe, the time evolution of the sphere radius takes the parametric form

$$r = A(1 - \cos \theta), \quad \text{and} \quad t = B(\theta - \sin \theta), \quad (5.269)$$

and the Newtonian equation of motion, $\ddot{r} = -G_N M/r^2$ implies that $A^3 = G_N M B^2$. The sphere collapses to a point ($r \rightarrow 0$) after a time $t_c = t|_{\theta=2\pi} = 2\pi B$. The mean density in the sphere is $\rho_{\text{int}} = 3M/4\pi r^3$, whereas that in the exterior space-time is $\rho_{\text{ext}} = (6\pi G_N t^2)^{-1}$ for an Einstein-de Sitter Universe. The ratio between these two densities gives the density contrast $\delta = \rho_{\text{int}}/\rho_{\text{ext}} - 1$.

At the beginning of the collapse, $\theta \ll 1$ so that

$$r(t) \simeq \frac{A}{2} \left(\frac{6t}{B} \right)^{2/3} \left[1 - \frac{1}{20} \left(\frac{6t}{B} \right)^{2/3} \right].$$

Initially, the sphere density contrast grows as $\delta \propto t^{2/3}$, in agreement with our previous analysis (5.22) in the linear regime.

This model shows the following stages of the gravitational collapse. If we compare the exterior evolution to the linear evolution in a flat Universe with $\Omega_m = 1$, three epochs can be distinguished.

1. *The decoupling of the expansion.* The overdensity is a system gravitationally bound. Its expansion reaches its maximum at $\theta = \pi$ and then decouples from the global expansion of the Universe. The density contrast then reaches the value $\delta = [A(6t/B)^{2/3}/2]/r^3 = 9\pi^2/16 \sim 5.55$ whereas the linear regime predicts $\delta = (3/20)(6\pi)^{2/3} \sim 1.06$.

2. *Gravitational collapse.* If no other mechanism comes into play, the sphere collapses into a singularity at $\theta = 2\pi$ and $\delta \rightarrow \infty$. The density contrast deduced from the linear analysis extrapolated until t_c is equal to $\delta = 1.69$.
3. *Virialization.* Some dissipative mechanisms convert the kinetic energy into thermal motion. The sphere then reaches a stationary regime and stabilizes its radius.

This basic analysis shows that the linear analysis is no longer valid as soon as $\delta > 1$. A more realistic approach to gravitational collapse in an expanding space-time is obtained from the solutions of the Einstein equations for a distribution of matter with spherical symmetry, known under the name of Lemaître–Tolman–Bondi [49] (see Chapter 3).

5.5.2.3 The non-linear power spectrum

A more precise analysis can be obtained by N -body numerical simulations. These simulations try to solve the Newtonian equations (5.15) and (5.16) and the Poisson equation (5.5). A detailed description of the various methods, problems and the state-of-the-art of these simulations can be obtained in Ref. [12]. Various codes [50] of N -body simulations are freely available.

From these simulations, various mappings have been established [51–53] between the matter power spectrum in the linear regime and the power spectrum in a non-linear regime. These mappings make it possible to discuss qualitatively the deformation of the power spectrum in the non-linear regime.

A fruitful hypothesis is that of *stable clustering* that postulates [54] that, in the non-linear regime, the regions of very high densities virialize, while keeping the mean density fixed. The correlation function of the density of these systems would then evolve as $\xi(r, t) \propto 1/\bar{\rho} \propto a^3$. It can then be shown [54] that, in a flat Universe, if the initial spectrum is of the form $P \propto k^n$, then this hypothesis implies that

$$\xi_{\text{nl}}(r, t) \propto r^{-\gamma}, \quad \text{with} \quad \gamma = \frac{3(3+n)}{5+n}. \quad (5.270)$$

The spectrum in the non-linear regime would then keep a memory of the initial spectrum.

We can thus postulate that the non-linear effects can be described by a universal mapping f_{nl} . Indeed, the correlation function, averaged on a large volume,

$$\bar{\xi}_{\text{nl}}(x) = \frac{3}{x^3} \int_0^x y^2 \xi_{\text{nl}}(y) dy,$$

measured by numerical simulations can be parameterized by a function of the linear regime correlation function [51]. The scale transformation between both regimes can be deduced by considering a collapsing spherical halo. This halo is constituted of spherical shells each of them reaching a maximal expansion before collapsing. If there is no shellcrossing, the mass in the interior of each one of them is conserved, so that

$$m(< r) = \frac{4\pi}{3} r^3 \rho(< r) = \frac{4\pi}{3} \ell^3 \rho(< \ell) = m_0(< \ell),$$

where $m_0(< \ell)$ is the initial mass distribution in a sphere of radius ℓ (in the linear regime). If we identify $1 + \bar{\xi}_{\text{nl}}$ with the density amplification factor, then

$$x^3 [1 + \bar{\xi}_{\text{nl}}(x, t)] = \ell^3,$$

where x represents the non-linear scale and ℓ the associated one in the linear regime. After this scale transformation, we suppose that $\bar{\xi}_{\text{nl}}(x, t) = f_{\text{nl}}[\bar{\xi}_{\text{lin}}(\ell, t)]$.

This procedure transposes easily into Fourier space. If we define

$$\Delta^2(k) = 4\pi k^3 P(k), \quad (5.271)$$

then the function f_{nl} is given by

$$\Delta_{\text{nl}}^2(k_{\text{nl}}) = f_{\text{nl}}[\Delta_{\text{lin}}^2(k)], \quad \text{with} \quad k^3 = \frac{k_{\text{nl}}^3}{1 + \Delta_{\text{nl}}^2(k_{\text{nl}})}. \quad (5.272)$$

Under the hypothesis of stable clustering, one can show [51] that this function must behave asymptotically, for $x \gg 1$, as

$$f_{\text{nl}}(x) \propto g^{-3}(\Omega) x^{3/2}, \quad g(\Omega) = \frac{D_+(a)}{a}, \quad (5.273)$$

where D_+ is the growing mode (5.22) of the linear regime. An analytical form of f_{nl} has been proposed; it is calibrated on N -body numerical simulations for Λ CDM models [51–53]. The existence of such a universal function can also be justified theoretically [55].

Figure 5.20 represents the result of numerical simulations, while Fig. 5.19 compares this analytical mapping to the results from numerical simulations and its effect on the power spectrum. These mappings will be useful when discussing the gravitational lensing by the large structures in Chapter 7.

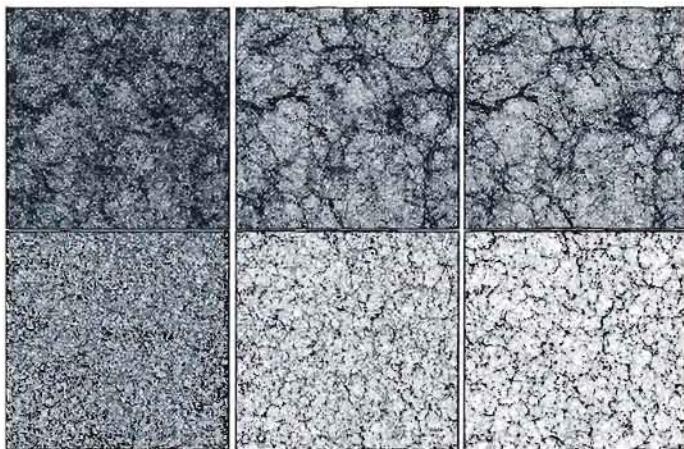


Fig. 5.20 Results of numerical simulations (top) for $n = -2$ to $1/(1+z) = 0.2, 0.45$ and 0.55 from right to left and (bottom) for $n = 0$ to $1/(1+z) = 0.25, 0.63$ and 0.83 . The influence of the spectrum on the morphology of the large structures (more filaments for $n = -2$ and smaller structures for $n = 0$) can be seen with the eye. From Ref. [53].

References

- [1] P.J.E. PEEBLES, 'Large-scale background temperature and mass fluctuations due to scale invariant primeval perturbations', *Astrophys. J.* **263**, L1, 1982.
- [2] G.R. BLUMENTHAL *et al.*, 'Formation of galaxies and large-scale structures with cold dark matter', *Nature* **311**, 517, 1984.
- [3] M. DAVIS *et al.*, 'The end of cold dark matter?', *Nature* **358**, 480, 1992.
- [4] S. WEINBERG, *Gravitation and cosmology: principles and applications of the general theory of relativity*, John Wiley and Sons, 1972.
- [5] T. PADMANABHAN, *Structure formation in the Universe*, Cambridge University Press, 1993.
- [6] S. DODELSON, *Modern cosmology*, Academic Press, 2003.
- [7] P.J.E. PEEBLES, *The large-scale structure of the Universe*, Princeton University Press, 1980.
- [8] J.A. PEACOCK, *Cosmological Physics*, Cambridge University Press, 1999.
- [9] A.R. LIDDLE and D.H. LYTH, *Cosmological inflation and large-scale structure*, Cambridge University Press, 2000.
- [10] L. LANDAU and E. LIFCHITZ, *Fluid mechanics*, MIR, 1964.
- [11] T. BUCHERT, 'A class of solutions in Newtonian cosmology and the pancake theory', *Astron. Astrophys.* **223**, 9, 1989.
- [12] F. BERNARDEAU *et al.*, 'Large-scale structure of the Universe and cosmological perturbation theory', *Phys. Rep.* **367**, 1, 2002.
- [13] O. LAHAV *et al.*, 'Dynamical effects of the cosmological constant', *Month. Not. R. Astron. Soc.* **251**, 128, 1991.
- [14] F. BERNARDEAU, *Cosmologie*, EDP Sciences, Paris, 2007.
- [15] A. DEKEL, *large-scale flows and cosmological implications*, in *Formation of structure in the Universe*, A. Dekel and J. Ostriker (eds.), Cambridge University Press, 1999
- [16] F. BERNARDEAU and J.-P. UZAN, 'Finite volume effects for non-Gaussian multi-field inflationary models', *Phys. Rev. D* **70**, 043533, 2004.
- [17] F. SYLOS-LABINI, M. MONTUORI and L. PIETRONERO, 'Scale-invariance of galaxy clustering', *Phys. Rep.* **293**, 61, 1998.
- [18] N. KAISER, 'On the spatial correlations of Abell clusters', *Astrophys. J.* **284**, L9, 1984.
- [19] J.M. BARDEEN *et al.*, 'The statistics of peaks of Gaussian random fields', *Astrophys. J.* **304**, 15, 1986.
- [20] J. FRY, 'The galaxy correlations hierarchy in perturbation theory', *Astrophys. J.* **279**, 499, 1984.
- [21] F. BOUCHET *et al.*, 'Weakly non-linear gravitational instability for arbitrary Ω ', *Astrophys. J.* **394**, L5, 1992.

- [22] F. BERNARDEAU, 'Skewness and kurtosis in large-scale cosmic fields', *Astrophys. J.* **433**, 1, 1994.
- [23] F. BERNARDEAU, 'The effects of smoothing on the statistical properties of large-scale cosmic fields', *Astron. Astrophys.* **291**, 697, 1994.
- [24] J.M. BARDEEN, 'Gauge invariant cosmological perturbations', *Phys. Rev. D* **22**, 1882, 1981.
- [25] U. GERLACH and U. SENGUPTA, 'Relativistic equations for aspherical gravitational collapse', *Phys. Rev. D* **18**, 1789, 1978.
- [26] H. KODAMA and M. SASAKI, 'Cosmological perturbation theory', *Prog. Theor. Phys. Suppl.* **78**, 1, 1984.
- [27] V.F. MUKHANOV, H.A. FELDMAN and R.H. BRANDENBERGER, 'Theory of cosmological perturbations', *Phys. Rep.* **5**, 203, 1992.
- [28] A.R. LIDDLE and D.H. LYTH, 'The cold dark matter density perturbation', *Phys. Rep.* **231**, 1, 1993.
- [29] R. DURRER, 'Gauge invariant cosmological perturbation theory', *Fund. Cosm. Phys.* **15**, 209, 1994.
- [30] S. HAWKING, 'Perturbations of an expanding Universe', *Astrophys. J.* **145**, 544, 1966.
- [31] G.F.R. ELLIS and M. BRUNI, 'Covariant and gauge-invariant approach to cosmological density fluctuations', *Phys. Rev. D* **40**, 1804, 1989.
- [32] G.F.R. ELLIS, *Relativistic cosmology*, Proc. Intl. School of Physics Enrico Fermi, Corso XLVII, R.K. Sachs (ed.) (Academic Press, 1971).
- [33] J.M. STEWART and M. WALKER, 'Perturbations of space-times in general relativity', *Proc. R. Soc. London A* **341**, 49, 1974.
- [34] J. MARTIN and P. PETER, 'On the causality argument in bouncing cosmologies', *Phys. Rev. Lett.* **92**, 061301, 2004.
- [35] J.M. BARDEEN, P.J. STEINHARDT and M.S. TURNER, 'Spontaneous creation of almost scale free density perturbations in an inflationary Universe', *Phys. Rev. D* **28**, 679, 1983.
- [36] D.H. LYTH, 'Large-scale energy-density perturbations and inflation', *Phys. Rev. D* **31**, 1792, 1985.
- [37] D. WANDS, K.A. MALIK, D.H. LYTH and A.R. LIDDLE, 'New approach to the evolution of cosmological perturbations on large scales', *Phys. Rev. D* **62**, 043527, 2000.
- [38] L. LANDAU, *Soviet J. Phys. (JETP)* **10**, 116, 1946.
- [39] C.P. MA and E. BERTSCHINGER, 'Cosmological perturbation theory in the synchronous and conformal gauge', *Astrophys. J.* **455**, 7, 1995.
- [40] P. MÉSZÁROS, 'The behaviour of point masses in an expanding cosmological substratum', *Astron. Astrophys.* **37**, 225, 1974.
- [41] J.-P. UZAN, 'Dynamics of relativistic interacting gases: from a kinetic to a fluid description', *Class. Quant. Grav.* **15**, 1063, 1998.
- [42] P.J.E. PEEBLES, *Principles of physical cosmology*, (Princeton University Press, 1993).
- [43] W. HU and N. SUGIYAMA, 'Anisotropies in the cosmic microwave background. an analytic approach', *Astrophys. J.* **444**, 489, 1995.

- [44] J. LESGOURGUÉS and S. PASTOR 'Massive neutrinos and cosmology', *Phys. Rep.* **49**, 307, 2006.
- [45] SDSS: <http://www.sdss.org>.
- [46] 2dFGRS: <http://www.ao.gov.au/2df/>.
- [47] M. TEGMARK , A. HAMILTON and Y. HU, 'The power spectrum of galaxies in 2dF 100k redshift survey', *Month. Not. R. Astron. Soc.* **335**, 887, 2002.
- [48] D. EISENSTEIN *et al.*, 'Detection of baryon acoustic peak in the large-scale correlation function of SDSS luminous red galaxies', *Astrophys. J.* **633**, 560, 2005.
- [49] R.C. TOLMAN, *Proc. Natl. Acad. Sci.* **20**, 169, 1934.
- [50] GADGET: Code in C described in V. SPRINGEL *et al.*, 'GADGET: A code for collisionless and gas dynamical cosmological simulations', *New. Astron.* **6**, 79, 2001, available at
<http://www.mpa-garching.mpg.de/gadget/right.html>;HYDRA: described in H. COUCHMAN *et al.*, 'Hydra Code Release', *Astrophys. J.* **452**, 797, 1995, available at
<http://coho.mcmaster.ca/hydra/hydra/hydra.html>;TPM: described in P. BODE *et al.*, 'The Tree-Particle-Mesh N-body Gravity Solver', *Astrophys. J. Suppl.* **128**, 561, 2000, available at
<http://www.astro.princeton.edu/~bode/TPM/>.
- [51] A. HAMILTON *et al.*, 'Reconstructing the primordial spectrum of fluctuations of the Universe from the observed nonlinear clustering of galaxies', *Astrophys. J. Lett.* **374**, L1, 1991.
- [52] J.A. PEACOCK and S.J. DODDS, 'Non-linear evolution of cosmological power spectra', *Month. Not. R. Astron. Soc.* **280**, L19, 1996.
- [53] R.E. SMITH *et al.*, 'Stable clustering, the halo model and non-linear cosmological power spectra', *Month. Not. R. Astron. Soc.* **341**, 1311, 2003.
- [54] P.J.E. PERRIERES, 'The gravitational instability picture and the nature of the distribution of galaxies', *Astrophys. J.* **189**, L51, 1974.
- [55] J.S. BAGLA and T. PADMANABHAN, 'Critical index and fixed point in the transfer of power in nonlinear gravitational clustering', *Month. Not. R. Astron. Soc.*, **286**, 1023, 1997.

6

The cosmic microwave background

The aim of this chapter is to provide the basis for the current understanding of the physics of the cosmic microwave background (CMB) and to detail the computation of the expected signatures from various scenarios.

We start by presenting in Section 6.1 a simplified version of the computation of the angular power spectrum of the temperature anisotropies in order to discuss the different contributions to this spectrum. The origin of the temperature anisotropies of the cosmic microwave background as well as the relationship between the temperature fluctuations observed today and those at decoupling are given in detail. This will allow us to understand and discuss in Section 6.2 the shape of the angular power spectrum.

We then present the modern description of the microwave background theory in Section 6.3. This description relies on a kinetic approach, which will require the study of the perturbed version of the Boltzmann equation in an expanding space-time.

Finally, we discuss the effects of various parameters on the angular power spectrum in Section 6.5, for both cosmological parameters (e.g. the different density contrasts, the Hubble constant, etc.) and parameters describing primordial physics.

For general reviews on this topic, see Refs. [1–4].

6.1 Origin of the cosmic microwave background anisotropies

After decoupling, photons no longer interact with matter and one can assume, to a first approximation, that they follow null geodesics. A simplified approach to the cosmic microwave background physics can then be obtained by computing the total redshift (between decoupling and today) of a photon propagating in a perturbed cosmological space-time.

In this simplified approach, we adopt a fluid description of matter and assume that the surface of last scattering is infinitely thin, i.e. we assume that decoupling happened instantaneously, (see Section 4.4.1, Chapter 4). This simplified presentation has the advantage of keeping the physics transparent and a more rigorous formalism, based on the kinetic approach, will be presented later in Section 6.3.

6.1.1 Sachs–Wolfe formula

The Sachs–Wolfe formula relates the present energy of a photon to its energy at decoupling [5, 6].

6.1.1.1 Propagation of a photon in a perturbed space-time

Let us start by studying the propagation of a photon in a space-time with a metric $g_{\mu\nu}$ of the form (5.52). The geodesic of this photon is a worldline $x^\mu(\lambda)$ where λ is an affine parameter. The tangent vector, k^μ , to this worldline satisfies the two equations (see Chapter 1)

$$k^\mu k_\mu = 0, \quad k^\mu \nabla_\mu k^\nu = 0.$$

Notice that two conformal space-times have the same causal structure so that their null geodesics will be identical (since these geodesics define the light cones). Defining the metric $\hat{g}_{\mu\nu}$ through $g_{\mu\nu} = a^2 \hat{g}_{\mu\nu}$, it can be easily checked that if k^μ is the tangent vector of a null geodesic of the metric g , then $\hat{k}^\mu = a^2 k^\mu$ will be the tangent vector of a null geodesic of the metric \hat{g} , i.e.

$$\hat{g}_{\mu\nu} \hat{k}^\mu \hat{k}^\nu = 0, \quad \hat{k}^\mu \hat{\nabla}_\mu \hat{k}^\nu = 0,$$

where $\hat{\nabla}_\mu$ is the covariant derivative associated with $\hat{g}_{\mu\nu}$.

Let us decompose the vector \hat{k}^μ as

$$\hat{k}^\mu = E (1 + M, \hat{e}^i + \delta \hat{e}^i). \quad (6.1)$$

E is a constant¹ and the four components of $\delta \hat{k}^\mu$ reduce to $(M, \delta \hat{e}^i)$. Only three of these four components are independent, since $\hat{k}^\mu \hat{k}_\mu = 0$ implies that

$$\gamma_{ij} \hat{e}^i \hat{e}^j = 1, \quad \hat{e}_j \delta \hat{e}^j = A + M - B_i \hat{e}^i - \frac{1}{2} h_{ij} \hat{e}^i \hat{e}^j. \quad (6.2)$$

To first order in perturbations, the geodesic equation for \hat{k}^μ takes the simplified form

$$\frac{d\delta \hat{k}^\nu}{ds} \equiv \hat{k}^\mu \partial_\mu \delta \hat{k}^\nu = -\delta \hat{\Gamma}_{\mu\rho}^\nu \hat{k}^\mu \hat{k}^\rho. \quad (6.3)$$

Using the decomposition (6.1), its time-like component ($\nu = 0$) gives

$$\frac{dM}{ds} = -A' - 2\hat{e}^i \partial_i A - \frac{1}{2} h'_{ij} \hat{e}^i \hat{e}^j + D_i B_j \hat{e}^i \hat{e}^j, \quad (6.4)$$

where we have set $dM/ds \equiv \hat{e}^i \partial_i M + \partial_0 M$.

6.1.1.2 Sachs-Wolfe formula

As seen in Chapter 1, see (1.122), the frequency, or equivalently the energy, of a photon measured by an observer comoving with the velocity field u^μ is related to its frequency at the time of emission by

$$\frac{E_0(e)}{E_E(x_E, \eta_E)} = \frac{(k^\mu u_\mu)_0}{(k^\mu u_\mu)_E}, \quad (6.5)$$

¹ As a simple check, remember that in a Friedmann-Lemaître space-time, $u_0 \propto a$ so that the energy of a photon scales as $k^\mu u_\mu = k^0 u_0 = (k^0/a^2)a$, that is as $1/a$ since $E = k^0$ is constant.

where u^μ is the baryon velocity field. The emission and reception points are related by

$$\mathbf{x}_E = \mathbf{x}_0 + e(\eta_0 - \eta_E), \quad (6.6)$$

where $e = -\hat{e}$ stands for the direction of observation. This trajectory is a geodesic in the unperturbed space-time. In what follows, we will work in the Born approximation in which all perturbed quantities are evaluated along an unperturbed trajectory.

We infer from (6.5) that the temperature observed in a direction e is related to the temperature at emission by

$$\frac{T_0(e)}{T_E(\mathbf{x}_E, \eta_E)} = \frac{a(\eta_E)}{a(\eta_0)} \left\{ 1 + [M + A + e^i(v_{bi} + B_i)]_E^0 \right\}, \quad (6.7)$$

where we have used $u_\mu = (-1 - A, v_k + B_k)$ and v_{bi} is the velocity perturbation of the baryon fluid (see Chapter 5).

Since recombination is dictated by the photon-electron interaction, which is itself governed by the coefficient $\sigma_T n_e$ (see Chapter 4, Section 4.4.1), we can model the surface of last scattering as a surface of constant electron density, $n_e = \text{const}$. Because of the density fluctuations, recombination does not occur at the same time at each point in the Universe (see Fig. 6.1). The time of emission is then decomposed as

$$\eta_E = \bar{\eta}_E + \delta\eta_E.$$

Both temperatures can be decomposed, respectively, as

$$T_E(\mathbf{x}_E, \eta_E) = \bar{T}_E(\eta_E) [1 + \Theta_E(\mathbf{x}_E, \eta_E)],$$

where \bar{T}_E is the spatial average of the temperature field at the time η_E , and

$$T_0(e) = \bar{T}_0(\eta_0) [1 + \Theta_0(e)].$$

$\Theta_0(e)$ represents the density fluctuations in a direction e compared to the average of the temperature on the sky, $\bar{T}_0(\eta_0)$.

Inserting these two decompositions into (6.7), we obtain that at lowest order in the perturbations

$$\frac{\bar{T}_0}{\bar{T}_E} = \frac{a(\bar{\eta}_E)}{a(\eta_0)}, \quad (6.8)$$

which is simply the scaling of the temperature of a blackbody in terms of the redshift in a homogeneous and isotropic expanding space-time [see Chapter 4].

Using that, to first order in the perturbations, $\bar{T}_E(\eta_E)a(\eta_E) = \bar{T}_E(\bar{\eta}_E)[1 + (T'/T)\delta\eta_E]a(\bar{\eta}_E)[1 + (a'/a)\delta\eta_E] = \bar{T}_E(\bar{\eta}_E)a(\bar{\eta}_E)$, we infer that the temperature contrast observed in the direction e is given by

$$\Theta_0(e) = \Theta_E(\mathbf{x}_E, \bar{\eta}_E) + [M + A + e^i(v_{bi} + B_i)]_E^0. \quad (6.9)$$

Again, in the Born approximation, the term in brackets being first order in the perturbations, it can be evaluated either in $\bar{\eta}_E$ or $\bar{\eta}_R$.

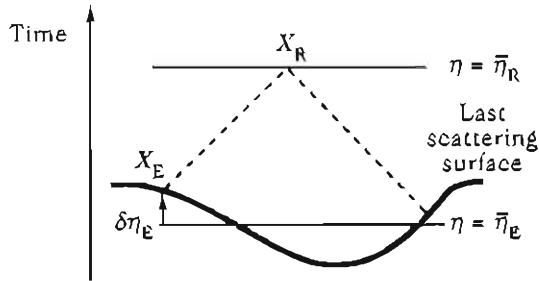


Fig. 6.1 The last-scattering surface is a surface of constant electron number density, $n_e = \text{const}$. Thus, it is not a constant-time hypersurface. The time of emission is related to the mean time by $\eta_E = \bar{\eta}_E + \delta\eta_E$.

To compute the first term on the right-hand side of the previous equation, we notice that the electron density, n_e , is a function of the photon and baryon energy densities, ρ_γ and ρ_b . As can be seen from Saha's equation [6], the effect of the baryon can be neglected, so that the last-scattering surface is well approximated by the surface $\rho_\gamma = \text{const}$. (see Chapter 4). Stefan-Boltzmann's law then implies that

$$\Theta_E(x_E, \bar{\eta}_E) = \frac{1}{4} \delta_\gamma(x_E, \bar{\eta}_E). \quad (6.10)$$

To evaluate the contribution $[M]_E^0$, we use the geodesic equation (6.4) that can be rewritten in the integral form

$$[M]_E^0 = -2[A]_E^0 + \int_E^0 \left(A' - \frac{1}{2} h'_{ij} e^i e^j + D_i B_j e^i e^j \right) ds, \quad (6.11)$$

the integral being performed along the unperturbed geodesic.

We infer from (6.9), (6.10) and (6.11) that

$$\begin{aligned} \Theta_0(e) = & \left[\frac{1}{4} \delta_\gamma + A - e^i (v_{bi} + B_i) \right] (x_E, \bar{\eta}_E) + \int_E^0 [A' - C' \right. \\ & \left. - e^i e^j (D_i D_j E' + D_i \bar{E}'_j - D_i B_j + \bar{E}'_{ij})] ds + f(0), \end{aligned} \quad (6.12)$$

where $f(0)$ is a function of the perturbation variables evaluated today. In terms of gauge invariant quantities, after expressing $e^j D_j (\bar{E}'_i - \bar{B}_i) = e^j D_j \bar{\Phi}_i$ as $-\bar{\Phi}'_i + d\bar{\Phi}_i/ds$, this expression takes the form

$$\begin{aligned} \Theta(x_0, \eta_0, e) = & \left[\frac{1}{4} \delta_\gamma^N + \bar{\Phi} - e^i (D_i V_b + \bar{V}_{bi} + \bar{\Phi}_i) \right] (x_E, \bar{\eta}_E) \\ & + \int_E^0 (\bar{\Phi}' + \Psi') [x(\eta), \eta] d\eta - \int_E^0 e^i \bar{\Phi}'_i [x(\eta), \eta] d\eta \\ & - \int_E^0 e^i e^j \bar{E}'_{ij} [x(\eta), \eta] d\eta. \end{aligned} \quad (6.13)$$

where we integrate along the unperturbed geodesic that we chose to parameterize as

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{e}(\eta_0 - \eta), \quad (6.14)$$

so that the integral now runs over the conformal time η . We have also voluntarily omitted the quantity $f(0)$ that is a function of the observation point and cannot be measured. This relation is known as the *Sachs–Wolfe equation*.

6.1.1.3 Interpretation

The relation (6.13) can be decomposed as the sum of three terms

$$\Theta_{\text{SW}} = \left(\frac{1}{4} \delta_{\gamma}^N + \Phi \right) (\mathbf{x}_E, \bar{\eta}_E) = \left(\frac{1}{4} \delta^F_{\gamma} + \Phi + \Psi \right) (\mathbf{x}_E, \bar{\eta}_E), \quad (6.15)$$

$$\Theta_{\text{dop}} = -e^i (V_{bi} + \bar{\Phi}_i) (\mathbf{x}_E, \bar{\eta}_E), \quad (6.16)$$

$$\Theta_{\text{ISW}} = \int_E^0 \left[(\Phi' + \Psi') - e^i \bar{\Phi}'_i - e^i e^j \bar{E}'_{ij} \right] d\eta. \quad (6.17)$$

The first term only has a scalar contribution and is called the *ordinary Sachs–Wolfe term*, the second has a scalar and vector contributions and represents the *Doppler term*, while the third one contains all three kinds of perturbations and is called the *integrated Sachs–Wolfe term*.

The ordinary Sachs–Wolfe term, Θ_{SW} , is evaluated at the decoupling time, i.e. in $(\mathbf{x}_E, \bar{\eta}_E)$, and contains two terms:

- the density contrast of the radiation fluid translates the fact that a denser region is hotter, by virtue of Stefan–Boltzmann's law,
- the gravitational potential indicates that a photon emitted in a potential well has an additional gravitational redshift (Einstein effect).

The term Θ_{dop} is simply a Doppler effect term, coming from the fact that the emitter and observer do not have the same velocity.

The integral term, Θ_{ISW} , depends on the photon's history between its emission and reception and contains the time derivative of the two gravitational potentials (in SVT decomposition). These terms will have a contribution coming from among other things,

- structures in formation traversed by the photon. Indeed, the gravitational potential of a collapsing structure, even spherical, is not ‘seen’ symmetrically by a photon that then acquires an additional redshift,
- in a Universe with a cosmological constant or non-vanishing curvature, we have seen that the gravitational potential is not constant in our recent past.

6.1.2 Angular power spectrum

The Sachs–Wolfe equation relates the observed temperature fluctuations to the cosmological perturbations. So $\Theta(\mathbf{x}_0, \eta_0, \mathbf{e})$ is a stochastic variable that can be characterized by its correlation function

$$C(\vartheta) = \langle \Theta(\mathbf{x}_0, \eta_0, \mathbf{e}_1) \Theta(\mathbf{x}_0, \eta_0, \mathbf{e}_2) \rangle. \quad (6.18)$$

The statistical isotropy, which follows from the cosmological principle, implies that this correlation function only depends on the relative angle between the two directions of observation, e_1 and e_2 ; $\cos \vartheta = e_1 \cdot e_2$. It is thus convenient to expand this correlation function in a basis of Legendre polynomials as

$$C(\vartheta) = \sum_{\ell} \frac{2\ell+1}{4\pi} C_{\ell} P_{\ell}(e_1 \cdot e_2), \quad (6.19)$$

which defines the angular power spectrum C_{ℓ} . A multipole ℓ corresponds approximately to an angular scale π/ϑ (see Appendix B) so that C_{ℓ} is a measure of the variance of the temperature fluctuations at this scale. If the temperature fluctuations have a Gaussian statistics, this function entirely characterizes the temperature distribution.

Expanding $\Theta(x_0, \eta_0, e)$ in spherical harmonics as

$$\Theta(x_0, \eta_0, e) = \sum_{\ell m} a_{\ell m}(x_0, \eta_0) Y_{\ell m}(e), \quad (6.20)$$

the expansion coefficients $a_{\ell m}(x_0, \eta_0)$ are then given by

$$a_{\ell m}(x_0, \eta_0) = \int d^2 e \Theta(x_0, \eta_0, e) Y_{\ell m}^*(e). \quad (6.21)$$

We can then show, using (B.17) to express the Legendre polynomial in (6.19) and then inverting this relation using (B.15) twice, that

$$(2\ell+1)C_{\ell} = \sum_m (a_{\ell m}(x_0, \eta_0) a_{\ell m}^*(x_0, \eta_0)). \quad (6.22)$$

We shall now compute this angular power spectrum as a function of the cosmological perturbations in the SVT decomposition. For each mode, we can expand the term $\Theta(x_0, \eta_0, e)$ in Fourier modes as

$$\Theta(x_0, \eta_0, e) = \int \frac{d^3 k}{(2\pi)^{3/2}} \hat{\Theta}(\eta_0, k, e), \quad (6.23)$$

where the phase $\exp(ik \cdot x_0)$ has been absorbed into the definition of $\hat{\Theta}$. We infer that

$$a_{\ell m}(x_0, \eta_0) = \int d^2 e \frac{d^3 k}{(2\pi)^{3/2}} \hat{\Theta}(\eta_0, k, e) Y_{\ell m}^*(e). \quad (6.24)$$

From now on, we omit the dependence of $a_{\ell m}$ on (x_0, η_0) .

6.1.2.1 Scalar modes

After expressing x_E with (6.6), the scalar part of (6.13), tells us that

$$\hat{\Theta}(\eta_0, k, e) = [\hat{\Theta}_{SW}(k) - ik\mu \hat{V}_b(k)] e^{ik\mu \Delta \eta e} + \int (\hat{\Phi}' + \hat{\Psi}') e^{ik\mu \Delta \eta} d\eta, \quad (6.25)$$

with $\Delta \eta = \eta_0 - \eta$ and where μ is defined by the relation $k \cdot e = k\mu$. The coefficient $\exp(ik\mu \Delta \eta)$ appears in the integral since the Fourier transform contains the

term $(\hat{\Phi}' + \hat{\Psi}') \exp[i\mathbf{k} \cdot \mathbf{x}(\eta)]$. The relation (6.14) then implies that $\exp[i\mathbf{k} \cdot \mathbf{x}(\eta)] = \exp(i\mathbf{k}\mu\Delta\eta) \exp(i\mathbf{k} \cdot \mathbf{x}_0)$ and we have assumed that the constant phase was absorbed into the definition of the Fourier coefficients [see (6.23)]. The same argument explains the coefficient $\exp(i\mathbf{k}\mu\Delta\eta_E)$ for the two first terms.

Each term of the expression (6.25) is a random variable that can be decomposed as $X(\mathbf{k}, \eta) = X(\mathbf{k}, \eta)a(\mathbf{k})$ where $a(\mathbf{k})$ is a random variable satisfying

$$\langle a(\mathbf{k})a^*(\mathbf{k}') \rangle = \delta^{(3)}(\mathbf{k} - \mathbf{k}'), \quad (6.26)$$

so that $\hat{\Theta}(\eta_0, \mathbf{k}, \mathbf{e}) = \hat{\Theta}(\eta_0, \mathbf{k}, \mu)a(\mathbf{k})$ with

$$\hat{\Theta}(\eta_0, \mathbf{k}, \mu) = [\hat{\Theta}_{SW}(\mathbf{k}) - \hat{V}_b(k)\partial_{\Delta\eta_E}] e^{ik\mu\Delta\eta_E} + \int (\hat{\Phi}' + \hat{\Psi}') e^{ik\mu\Delta\eta} d\eta. \quad (6.27)$$

Using the relation (B.21) to expand the exponential factor in terms of spherical harmonics, we obtain

$$\hat{\Theta}(\eta_0, \mathbf{k}, \mu) = 4\pi \sum_{\ell m} \hat{\Theta}_\ell^{(S)}(k) Y_{\ell m}^*(\hat{\mathbf{k}}) Y_{\ell m}(\mathbf{e}), \quad (6.28)$$

with

$$\hat{\Theta}_\ell^{(S)}(k) = \hat{\Theta}_{SW}(k) j_\ell(k\Delta\eta_E) - \frac{\hat{V}_b(k)}{k} j'_\ell(k\Delta\eta_E) + \int (\hat{\Phi}' + \hat{\Psi}') j_\ell(k\Delta\eta) d\eta. \quad (6.29)$$

Inserting these expressions into (6.22) and then using (6.26), we obtain

$$\begin{aligned} (2\ell + 1)C_\ell^S &= \sum_m \int d^2\mathbf{e}_1 d^2\mathbf{e}_2 \frac{d^3k}{(2\pi)^3} \Theta(\eta_0, \mathbf{k}, \mu_1) \Theta^*(\eta_0, \mathbf{k}, \mu_2) Y_{\ell m}(\mathbf{e}_1) Y_{\ell m}^*(\mathbf{e}_2), \\ &= \frac{2}{\pi} \int k^2 dk \sum_{\ell_1 m_1 \ell_2 m_2 m} \hat{\Theta}_{\ell_1}^{(S)}(k) \hat{\Theta}_{\ell_2}^{(S)*}(k) \int d^2\hat{\mathbf{k}} d^2\mathbf{e}_1 d^2\mathbf{e}_2 \\ &\quad Y_{\ell_1 m_1}(\hat{\mathbf{k}}) Y_{\ell_2 m_2}^*(\hat{\mathbf{k}}) Y_{\ell_1 m_1}(\mathbf{e}_1) Y_{\ell_2 m_2}^*(\mathbf{e}_2) Y_{\ell m}^*(\mathbf{e}_2). \end{aligned}$$

The integration over $\hat{\mathbf{k}} = \mathbf{k}/k$ leads to a term $\delta_{\ell_1 \ell_2} \delta_{m_1 m_2}$, and the integrations over \mathbf{e}_1 and \mathbf{e}_2 to $\delta_{\ell_1 \ell_2} \delta_{m_1 m_2}$ and $\delta_{\ell \ell_2} \delta_{m m_2}$, respectively, so that

$$C_\ell^S = \frac{2}{\pi} \int |\hat{\Theta}_\ell^{(S)}(k)|^2 k^2 dk. \quad (6.30)$$

Fig. 6.2 decomposes the power spectrum of the scalar modes into its ordinary Sachs-Wolfe, Doppler and integrated Sachs-Wolfe contributions.

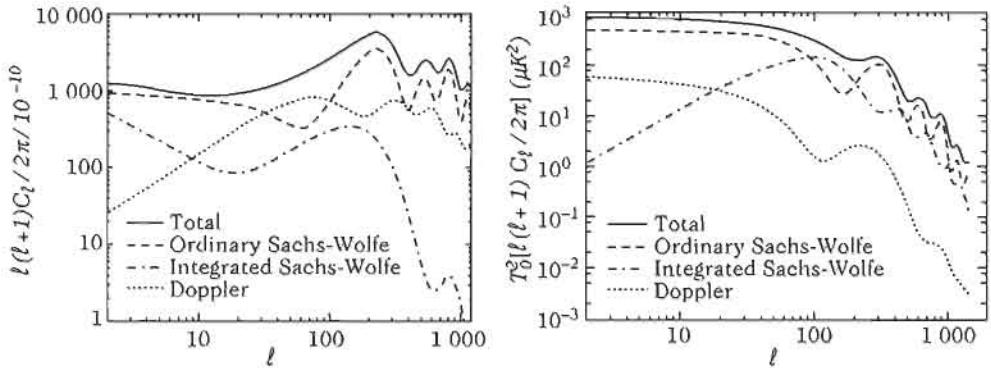


Fig. 6.2 The angular power spectrum and its three contributions, ordinary Sachs-Wolfe, Doppler and integrated Sachs-Wolfe for adiabatic (left) and isocurvature (right) initial conditions and a scale invariant primordial spectrum ($n_s = 1$), as predicted by inflation (see Chapter 8).

6.1.2.2 Vector modes

We now perform the same exercise for the vector contributions. The decomposition (6.23) now gives

$$\hat{\Theta}(\eta_0, \mathbf{k}, \mathbf{e}) = -e^i [\bar{V}_{bi}(\mathbf{k}) + \bar{\Phi}_i(\mathbf{k})] e^{ik\mu\Delta\eta_E} - \int e^i \bar{\Phi}'_i e^{ik\mu\Delta\eta} d\eta, \quad (6.31)$$

The vector variables are transverse, so that their Fourier coefficients do not have any components parallel to the wavevector. They can thus be expanded in the basis $(\mathbf{e}_1, \mathbf{e}_2)$ of the space perpendicular to \mathbf{k} as

$$\bar{X}_i(\mathbf{k}, \eta) = \sum_{\lambda=\pm} \bar{X}_{\lambda}(k, \eta) \mathbf{e}_i^{\lambda}(\mathbf{k}) a_{\lambda}(k), \quad (6.32)$$

with the complex quantities $\mathbf{e}_{\pm}^i = \mathbf{e}_1^i \pm i\mathbf{e}_2^i$ so that $e_{\pm}^i e_i = \sin \theta \exp(\pm i\varphi)$. Both polarizations are independent and the random variables $a_{\lambda}(k)$ now satisfy the relation

$$\langle a_{\lambda}(k) a_{\lambda'}^*(k') \rangle = \delta^{(3)}(\mathbf{k} - \mathbf{k}') \delta_{\lambda\lambda'}, \quad (6.33)$$

so that after using the relation (B.21) to expand the exponential factor in terms of spherical harmonics, we obtain

$$\hat{\Theta}(\eta_0, \mathbf{k}, \mathbf{e}) = 4\pi \sum_{\lambda=\pm} \sum_{\ell m} (\sin \vartheta e^{i\lambda\varphi}) \tilde{\Theta}_{\lambda,\ell}^{(V)}(k) Y_{\ell m}(k) Y_{\ell m}^*(\mathbf{e}) a_{\lambda}(k), \quad (6.34)$$

with

$$\tilde{\Theta}_{\lambda,\ell}^{(V)}(k) = -[\bar{V}_{\lambda}(k) + \bar{\Phi}_{\lambda}(k)] j_{\ell}(k\Delta\eta_E) - \int \bar{\Phi}'_{\lambda}(k) j_{\ell}(k\Delta\eta) d\eta. \quad (6.35)$$

Inserting these expressions in the relation (6.22), using the relation (6.33) and then performing the integral over $d^2 k$, we obtain

$$(2\ell+1)C_\ell^V = \frac{2}{\pi} \int k^2 dk \sum_{\ell_1, \lambda} \left| \tilde{\Theta}_{\lambda, \ell_1}^{(V)}(k) \right|^2 \sum_{mm_1} \left| \int d^2 e \sin \vartheta e^{i\lambda\varphi} Y_{\ell m}(e) Y_{\ell_1 m_1}^*(e) \right|^2.$$

In order to compute the integral over e , first notice that $\sin \vartheta e^{i\lambda\varphi} = -\lambda 2\sqrt{2\pi/3} Y_{1\lambda}(e)$ before using the integral (B.23). We can then perform the sum over m and m_1 using the relation (B.25) of the $3j$ -Wigner symbols. This sum only keeps the terms with $\ell_1 = \ell \pm 1$ so that the Bessel functions j_{ℓ_1} that appeared in the expression for $\Theta_{\lambda, \ell_1}^{(V)}$ recombine to lead to the function [see (B.46)]

$$j_\ell^{(11)}(x) = \sqrt{\frac{\ell(\ell+1)}{2}} \frac{j_\ell(x)}{x}. \quad (6.36)$$

We finally obtain

$$C_\ell^V = \frac{2}{\pi} \int \sum_\lambda \left| \tilde{\Theta}_{\lambda, \ell}^{(V)}(k) \right|^2 k^2 dk, \quad (6.37)$$

$$\hat{\Theta}_{\lambda, \ell}^{(V)}(k) = [\bar{V}_\lambda(k) + \bar{\Phi}_\lambda(k)] j_\ell^{(11)}(k\Delta\eta_E) + \int \bar{\Phi}'_\lambda(k) j_\ell^{(11)}(k\Delta\eta) d\eta. \quad (6.38)$$

6.1.2.3 Tensor modes

The computation for the tensor modes is noticeably the same as that for the vector modes. The expansion (6.23) now gives

$$\hat{\Theta}(\eta_0, k, e) = - \int e^j e^i \bar{E}'_{ij} e^{ik\mu\Delta\eta} d\eta. \quad (6.39)$$

The tensor \bar{E}_{ij} can be expanded as

$$\bar{E}_{ij}(k, \eta) = \sum_\lambda \bar{E}_\lambda(k, \eta) \varepsilon_{ij}^\lambda a_\lambda(k), \quad (6.40)$$

where the random variable $a_\lambda(k)$ satisfies the condition (6.33) and where the polarization tensor is defined as

$$\varepsilon_{ij}^\lambda = e_i^+ e_j^+ \delta_\lambda^+ + e_i^- e_j^- \delta_\lambda^-. \quad (6.41)$$

In particular, this implies that $\varepsilon_{ij}^\lambda e^i e^j = \sin^2 \theta \exp(2i\lambda\varphi)$.

After using (B.21) to expand the exponential factor, we find

$$\hat{\Theta}(\eta_0, k, e) = 4\pi \sum_{\lambda=\pm} \sum_{\ell m} (\sin^2 \vartheta e^{2i\lambda\varphi}) \tilde{\Theta}_{\lambda, \ell}^{(T)}(k) Y_{\ell m}(k) Y_{\ell m}^*(e) a_\lambda(k), \quad (6.42)$$

with

$$\tilde{\Theta}_{\lambda, \ell}^{(T)}(k) = - \int \bar{E}'_\lambda(k) j_\ell(k\Delta\eta) d\eta. \quad (6.43)$$

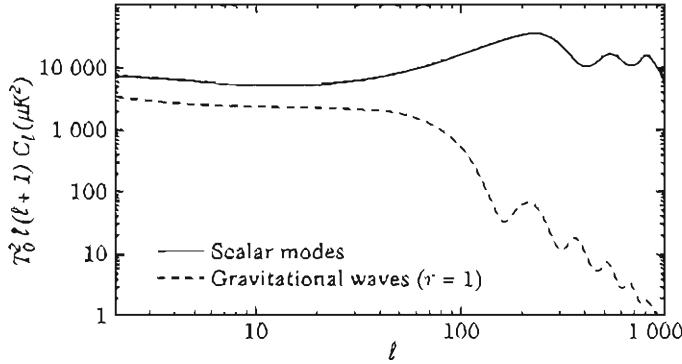


Fig. 6.3 Angular power spectrum for the tensor modes, compared to that of scalars for a ratio $r = 1$ between the primordial spectra of the two modes.

Inserting these expressions in (6.22), using (6.33) and then performing the integral over $d^2\vec{k}$ we find that

$$(2\ell + 1)C_\ell^T = \frac{2}{\pi} \int k^2 dk \sum_{\ell_1, \lambda} \left| \tilde{\Theta}_{\lambda, \ell_1}^{(T)}(k) \right|^2 \sum_{m m_1} \left| \int d^2 e \sin^2 \vartheta e^{2i\lambda\varphi} Y_{\ell m}(e) Y_{\ell_1 m_1}^*(e) \right|^2.$$

To compute the integral over e , the procedure is the same as for the vector modes, noticing that now $\sin^2 \vartheta e^{\pm 2i\varphi} = 4\sqrt{2\pi}/15 Y_{2\pm 2}(e)$. The sum over the $3j$ -Wigner symbols, gives the terms $\ell_1 = \ell \pm 2$ and $\ell_1 = \ell$. Using (B.47), the terms j_{ℓ_1} in the expression for $\Theta_{\lambda, \ell_1}^{(T)}$ combine, leading to functions [see (B.46)]

$$j_\ell^{(22)}(x) = \sqrt{\frac{(\ell + 2)!}{(\ell - 2)!}} \frac{j_\ell(x)}{x^2}. \quad (6.44)$$

We finally obtain

$$C_\ell^T = \frac{1}{\pi} \int \sum_\lambda \left| \hat{\Theta}_{\lambda, \ell}^{(T)}(k) \right|^2 k^2 dk, \quad (6.45)$$

$$\hat{\Theta}_{\lambda, \ell}^{(T)}(k) = - \int \bar{E}'_\lambda(k) j_\ell^{(22)}(k \Delta \eta) d\eta. \quad (6.46)$$

6.2 Properties of the angular power spectrum

The previous description allows us to understand the general shape of the power spectra of the scalar and tensor modes. The vector modes will be neglected in the rest of this chapter. Complementary references on the study of the properties of the temperature power spectrum can be found in Refs. [2, 7].

6.2.1 Large angular scales

For large angular scales, i.e. scales corresponding to super-Hubble modes at the time of decoupling, the gravitational potential remains constant (Chapter 5). For these modes, we can estimate the angular power spectrum at small ℓ as a function of the characteristics of the power spectrum of the gravitational potential. This part of the angular spectrum is usually called the *Sachs–Wolfe plateau*.

6.2.1.1 Adiabatic perturbations

The initial conditions (5.223) imply that initially $\delta_\gamma^N(0) = -2\Psi(0)$, i.e. that the photon overdense regions correspond to potential wells, and $V_\gamma = 0$. On super-Hubble scales at the time of decoupling, V_γ did not have time to grow significantly. Equation (5.255) then implies that $\delta_\gamma^N - 4\Psi$ remains constant. In the matter era, $\delta_\gamma^N/4 = \delta_m^N/3 = -2\Phi/3$ [(5.220) and (5.227)] so that

$$\Theta_{SW} = \frac{1}{3}\Phi \quad (6.47)$$

is the main contribution to the temperature fluctuations, in agreement with Fig. 6.2. Equation (6.30) then implies that the angular power spectrum is the convolution of the primordial spectrum with Bessel functions

$$C_\ell^S = \frac{2}{\pi} \int j_\ell^2[k(\eta_0 - \eta_{LSS})] \frac{1}{9} P_\Phi(k) k^3 \frac{dk}{k}, \quad (6.48)$$

since only the first term of (6.29) contributes significantly and $\Delta\eta_E = \eta_0 - \eta_{LSS}$, where $\eta_E = \eta_{LSS}$ is the time of last scattering. In the matter era, the gravitational potential is constant and is related to the curvature perturbation by $\Phi = 3\zeta/5$ for super-Hubble modes (5.149). It follows that $P_\Phi = 9P_\zeta/25$. Using the definition $\mathcal{P}_X = k^3 P_X / 2\pi^2$, we infer that

$$C_\ell^S = \pi \int A_S^2(k) j_\ell^2[k(\eta_0 - \eta_{LSS})] \frac{dk}{k}, \quad (6.49)$$

using a spectrum of the form (8.227) as predicted by inflation, with the definition (8.234) giving $A_S^2 = 4\mathcal{P}_\zeta/25$. This integral is dominated by modes such as $k\eta_0 \sim \ell$ so that one expects to have $C_\ell \propto A_S^2(k \sim \ell/\eta_0)$. More precisely, for a power-law spectrum (see Appendix B),

$$C_\ell^S \simeq \pi A_S^2(k_0) \left[\frac{\pi}{2} \frac{\Gamma(3 - n_S)}{2^{3-n_S}} \frac{\Gamma\left(\ell + \frac{n_S - 1}{2}\right)}{\Gamma^2\left(2 - \frac{n_S}{2}\right) \Gamma\left(\ell + \frac{5 - n_S}{2}\right)} \right], \quad (6.50)$$

choosing the pivot $k_* = k_0 = 1/\eta_0$, and with the approximation $\eta_{LSS} \ll \eta_0$. For $n_S = 1$, we obtain

$$\ell(\ell + 1)C_\ell^S = \frac{\pi}{2} A_S^2(k_0), \quad (6.51)$$

which corresponds to the plateau of the angular power spectrum at low multipoles (Fig. 6.2). For large ℓ , $\Gamma(\ell + a) \propto \ell^a$ so that

$$\ell(\ell+1)C_\ell^S \propto \ell^{n_S - 1}. \quad (6.52)$$

The spectral index of the primordial spectrum influences directly the slope of the Sachs-Wolfe plateau. Also notice that (6.47) implies that $\Theta_{SW} = -\delta_\gamma^N/6$. So the cold regions correspond to over-dense regions at the time of decoupling.

6.2.1.2 Isocurvature perturbations

For isocurvature perturbations, we initially have $\delta_\gamma^N(0) = \Psi(0) = 0$ and $V_\gamma(0) = 0$ in the radiation era (Chapter 5). If $S(0)$ is the initial entropy, then in the matter era $\Phi = -S/5$. For long-wavelength modes, $\delta_\gamma^N - 4\Psi$ remains constant so that $\delta_\gamma^N/4 + \Phi = \Psi + \Psi \simeq 2\Phi$. The leading term in the temperature fluctuation in the matter era is thus

$$\Theta_{SW} = 2\Phi = -2S/5; \quad (6.53)$$

see (5.229). Using the same argument as for the adiabatic modes, we obtain

$$C_\ell^S \simeq 36\pi A_S^2(k_0) \left[\frac{\pi}{2} \frac{\Gamma(3-n_S)}{2^{3-n_S}} \frac{\Gamma\left(\ell + \frac{n_S-1}{2}\right)}{\Gamma^2\left(2 - \frac{n_S}{2}\right) \Gamma\left(\ell + \frac{5-n_S}{2}\right)} \right]. \quad (6.54)$$

For $n_S = 1$, we have again a plateau at large angular scales

$$\ell(\ell+1)C_\ell^S = 36\pi A_S^2(k_0). \quad (6.55)$$

6.2.2 Intermediate scales

The origin and structure of the peaks of the angular power spectrum can be understood by studying the system of photons and baryons during the radiation era.

The evolution of the density of photons coupled to baryons is dictated by the equations (5.255) and (5.256), as well as (5.253) and (5.254)

$$\delta_\gamma^{N'} = \frac{4}{3}k^2V_\gamma + 4\Psi', \quad V'_\gamma = -\frac{1}{4}\delta_\gamma^N - \Phi + \frac{1}{6}k^2\pi_\gamma + \tau'(V_b - V_\gamma), \quad (6.56)$$

$$\delta_b^{N'} = k^2V_b + 3\Psi', \quad V'_b = -\mathcal{H}V_b - \Phi + \frac{\tau'}{R}(V_\gamma - V_b), \quad (6.57)$$

with $\tau' = an_e\sigma_T$. The residual anisotropic pressure term is related to the radiation velocity by

$$\frac{k^2}{12}\pi_\gamma = -\frac{8}{45}\frac{k^2}{\tau'}V_\gamma. \quad (6.58)$$

This relation reflects the fact that locally, the quadrupole can be generated from the gradient of a velocity field. It will be justified when the effect of polarization on Thomson scattering will be taken into account [see (6.231) below]. Notice that if the polarization of radiation is neglected, the factor 8/45 is replaced by a factor 2/15.

6.2.2.1 Dynamics of the photon–baryon fluid

Let us start by considering wavelengths that are large compared to $1/\tau'$. The Euler equations for the photons and the baryons can be expanded to leading order in k/τ' to obtain, if the anisotropic pressure term is neglected,

$$V_b = V_\gamma, \quad \delta'_\gamma - \frac{4}{3}\delta'_b = 0.$$

The second equation implies that the entropy remains constant during the evolution. In this tight coupling regime, we can combine both Euler equations to obtain

$$[(1+R)V_\gamma]' = -\frac{1}{4}\delta_\gamma^N - (1+R)\Phi. \quad (6.59)$$

The continuity equation for photons can then be used to deduce that the evolution of the density contrast is dictated by

$$\delta_\gamma^{NN} + \frac{R'}{1+R}\delta_\gamma^{N'} + k^2 c_s^2 \delta_\gamma^N = 4F(\Phi, \Psi) = 4\left(\Psi'' + \frac{R'}{1+R}\Psi' - \frac{1}{3}k^2\Phi\right), \quad (6.60)$$

with

$$R = \frac{3}{4}\frac{\rho_b}{\rho_\gamma}, \quad c_s^2 = \frac{1}{3}\frac{1}{1+R}. \quad (6.61)$$

This equation describes a forced damped oscillator. The forcing term is mostly determined by the Bardeen potentials, which tend to be mostly dominated by the density contrast of the CDM as R increases. It induces a shift mean value of the oscillations. Equation (6.60) can take the more compact form

$$[(1+R)\delta_\gamma^{N'}]' + \frac{1}{3}k^2\delta_\gamma^N = 4\left\{[(1+R)\Psi']' - \frac{1}{3}k^2(1+R)\Phi\right\}. \quad (6.62)$$

This equation is at the basis of the understanding of the acoustic oscillations of the photon–baryon plasma. The density contrast of the photons oscillates due to the two opposite forces: the radiation pressure and gravity. The frequency of these oscillations is $\omega_s = kc_s$. In a WKB regime where the period of the oscillations is small compared to the characteristic time of the variation of the amplitude [that is when $\delta_\gamma^{N'} \ll \delta_\gamma^N \omega_s$ and $\delta_\gamma^{NN} \ll (\delta_\gamma^N \omega_s)'$ but taking into account that $\omega_s'/\omega_s = -R'/2(1+R)$ which is of the same order as the damping term], the homogeneous equation has two solutions

$$\delta_a = (1+R)^{-1/4} \cos(kr_s), \quad \delta_b = (1+R)^{-1/4} \sin(kr_s), \quad (6.63)$$

where

$$r_s = \int_0^\eta c_s d\eta' = \int_0^\eta \frac{d\eta'}{\sqrt{3(1+R)}} \quad (6.64)$$

is the sound horizon at time η . The general solution of (6.60) is thus of the form

$$\delta_\gamma^N = C_a \delta_a(\eta) + C_b \delta_b(\eta) + 4 \int_0^\eta G(\eta, \eta') F(\eta') d\eta',$$

with the Green function

$$\mathcal{G}(\eta, \eta') \equiv \frac{\delta_a(\eta')\delta_b(\eta) - \delta_a(\eta)\delta_b(\eta')}{\delta_a(\eta')\delta'_b(\eta') - \delta'_a(\eta')\delta_b(\eta')}, = \frac{\sqrt{3}}{k} \frac{[1 + R(\eta')]^{3/4}}{[1 + R(\eta)]^{1/4}} \sin [kr_s(\eta) - kr_s(\eta')], \quad (6.65)$$

so that the general solution is

$$[1 + R(\eta)]^{1/4} \delta_\gamma^N = \delta_\gamma(0) \cos [kr_s(\eta)] + \frac{\sqrt{3}}{k} \left[\delta'_\gamma(0) + \frac{1}{4} R'(0) \delta_\gamma(0) \right] \sin [kr_s(\eta)] \\ + 4 \frac{\sqrt{3}}{k} \int_0^\eta [1 + R(\eta')]^{3/4} \sin [kr_s(\eta) - kr_s(\eta')] F(\eta') d\eta'. \quad (6.66)$$

6.2.2.2 Sachs-Wolfe term

Let us assume to start with that the Bardeen potentials are almost constant. This corresponds to a matter-dominated Universe. Furthermore, we shall neglect the time variation of R . The solution (6.66) then implies

$$\Theta_{SW} = [\Theta_{SW}(0) + R\Phi] \cos kr_s(\eta) + \frac{\Theta'_{SW}(0)}{kc_s} \sin kr_s(\eta) - R\Phi, \quad (6.67)$$

and r_s reduces to $r_s = c_s \eta$. The Sachs-Wolfe term undergoes oscillations in time that are the signature of the density waves, i.e. sound waves, propagating in the photon-baryon plasma with a velocity c_s . Notice that the minimum of the oscillations, $-R\Phi$, is shifted with respect to zero because of the baryons [compared with the study of the pure radiation fluid in Chapter 5]. The amplitude of the oscillations increases with R and thus with the amount of baryons.

In reality, the variation of R and of the Bardeen potentials cannot be neglected over several periods and the evolution of Θ_{SW} is obtained from (6.66). Its amplitude thus decreases as $(1 + R)^{1/4}$ during the cosmic expansion. Figure 6.4 illustrates the time evolution of Θ_{SW} for several modes before decoupling.

6.2.2.3 Doppler term

Neglecting the time dependence of the Bardeen potentials and of R , the continuity equation for the photons implies that $k^2 V_\gamma / 3 = \Theta'_{SW} + (\Psi - \Phi)',$ so that in this approximation,

$$\frac{kV_\gamma}{3} = -c_s |\Theta_{SW}(0) + R\Phi| \sin kr_s(\eta) + \frac{\Theta'_{SW}(0)}{k} \cos kr_s(\eta). \quad (6.68)$$

In the tight coupling limit, this velocity can be identified with that of the baryons that is at the origin of the Doppler contribution. Since the direction of the baryons is random, we can estimate that $\Theta_{Dop} \sim kV_\gamma/\sqrt{3}$. This term oscillates symmetrically around zero, in quadrature with Θ_{SW} , with an amplitude $\sqrt{3}c_s = (1 + R)^{-1/2}$ smaller. The relation between the phases of the density and the velocity implies that the peaks associated with the Doppler term fill in the troughs of the spectrum of the Sachs-Wolfe term. When $R \rightarrow 0$, the Doppler term has the same amplitude as the Sachs-Wolfe term, so that the resulting oscillations of the spectrum are very damped. When R becomes large enough the Doppler term is smaller and the final spectrum then has oscillations, but less marked than that of the spectrum of Θ_{SW} (Fig. 6.2).

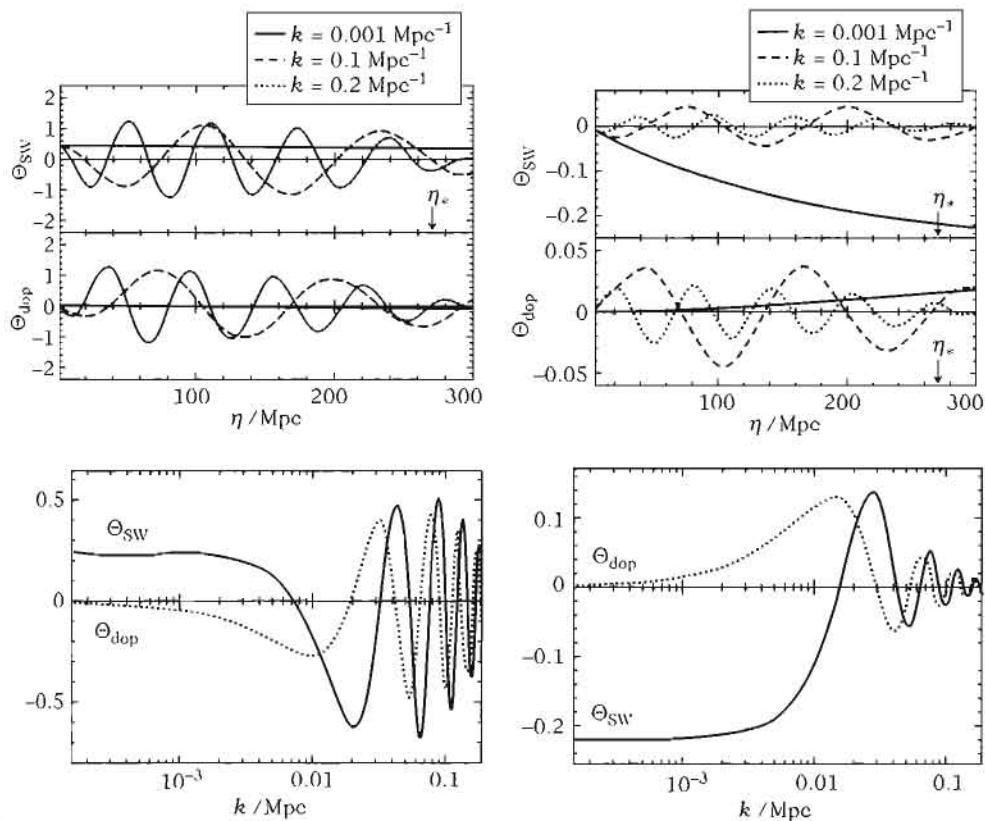


Fig. 6.4 Evolution of $\Theta_{\text{SW}} = \delta_\gamma^N/4 + \Phi$ and of $\Theta_{\text{dop}} \sim kV_\gamma/\sqrt{3}$ as a function of time for three modes with wave number $k = 10^{-3}$ (solid line), 10^{-1} (dashed) and $2 \times 10^{-1} \text{ Mpc}^{-1}$ (dotted). The contribution at the time of decoupling as a function of k is presented at the bottom. We compare the adiabatic (left) to the isocurvature (right) initial conditions.

6.2.2.4 Positions of the acoustic peaks

The position of the acoustic peaks depends on the initial conditions of the perturbations. Below the sound horizon, the Bardeen potentials are negligible. As will be seen in the next section, as soon as the scales become of the same order as the scattering scale (6.80), the photon density contrast is exponentially damped. We then obtain

$$\Theta_{\text{SW}} = C_A \frac{\mathcal{D}(k, \eta)}{(1+R)^{1/4}} \cos kr_s + C_I \frac{\mathcal{D}(k, \eta)}{(1+R)^{1/4}} \sin kr_s, \quad (6.69)$$

with $\mathcal{D}(k, \eta) \equiv \exp(-k^2/k_D^2)$. C_A and C_I are two integration constants, the values of which depend on the initial conditions.

- *Adiabatic initial conditions:* in this case, initially $\delta_\gamma^N(0) = -2\Phi(0)$. If the photon velocity V_γ is negligible, the conservation equation reduces to $\delta_\gamma^{N'} = 4\Phi'$, so that

$\delta_\gamma^N(\eta) = -2\Phi(0) + 4\Phi(\eta) - 4\Phi(0)$. After the Hubble crossing, the gravitational potential tends to zero so that $\delta_\gamma^N(\eta) \simeq -6\Phi(0)$. So, at the time where the mode enters the sound horizon $\delta_\gamma^N/4 = -3\Phi(0)/2$ and $\delta_\gamma^{N'} = 0$, so that $C_A = -3\Phi(0)/2$ and $C_I = 0$,

$$\Theta_{SW} = -\frac{3}{2}\Phi(0)\frac{\mathcal{D}(k, \eta)}{(1+R)^{1/4}} \cos kr_s. \quad (6.70)$$

For adiabatic perturbations, the gravitational potential is constant so that F is constant in the (6.60). In the limit $kr_s \rightarrow 0$, this constant excites the cosine mode.

- *Isocurvature initial conditions:* initially $\delta_\gamma^N(0) = \Phi(0) = 0$. As can be seen from (5.220), $\delta_\gamma^N \propto -aS$ in the radiation era and $S = S(0)$ remains constant. The gravitational potential then grows since $\Phi \propto -aS$ (5.228). During this era $a \propto \eta$ so that $F \propto \eta$ in (6.60). In the limit $kr_s \rightarrow 0$, this term excites the sine mode. Solving the evolution equations, we then obtain

$$C_A = 0, \quad C_I = \frac{\sqrt{6}}{4} \frac{k_{eq}}{k} S(0). \quad (6.71)$$

The factor k_{eq}/k can be understood as $\delta_\gamma^N \propto -aS$ until the mode reaches the Jeans length, i.e. until $a \sim k_{eq}/k$.

The term in $\cos(kr_s)$ or $\sin(kr_s)$ is excited, depending on the nature of the initial conditions. We infer that at the time of decoupling ($\eta = \eta_{LSS}$), $\Theta_{SW}(k, \eta_{LSS})$ is maximal for

$$k_{(p)}r_s(\eta_{LSS}) = \begin{cases} p\pi, & (\text{adiabatic}) \\ \left(p - \frac{1}{2}\right)\pi, & (\text{isocurvature}) \end{cases} \quad (6.72)$$

with $p = 1, 2, \dots$. The physical scale corresponding to these wave numbers, $\lambda_{(p)} = a(\eta_{LSS})\pi/k_{(p)}$, is observed with an angular scale of $\vartheta_{(p)} \sim \lambda_{(p)}/D_A(\eta_{LSS})$, where D_A is the angular distance (3.74), i.e.

$$\vartheta_{(p)} = \frac{\pi}{k_{(p)}f_K(z_{LSS})}. \quad (6.73)$$

An angular scale ϑ corresponds approximatively to a multipole $\ell \sim \pi/\vartheta$ (see the properties of the Legendre polynomials, Appendix A). The angular power spectrum therefore has a set of peaks located at the multipoles

$$\ell_{(p)} = \frac{f_K(z_{LSS})}{r_s(z_{LSS})} \times \begin{cases} p, & (\text{adiabatic}) \\ p - \frac{1}{2}, & (\text{isocurvature}). \end{cases} \quad (6.74)$$

The position of the peaks thus mainly depends on the cosmological parameters via the values of z_{LSS} and $r_s(z_{LSS})$, and very little on the primordial spectrum of the perturbations. Adiabatic models have a set of peaks from the series (1:2:3:...), while

the peaks series of the isocurvature models is (1:3:5:...). The peak $p = 1$ in the isocurvature models is usually too weak so that the first peak is localized at

$$\ell_{(1)} \sim \begin{cases} 220, & (\text{adiabatic}) \\ 330, & (\text{isocurvature}) \end{cases} \quad (6.75)$$

for a flat Universe with no cosmological constant. This difference in the peaks position makes it possible to constrain the relative importance of the two types of perturbations. The angular spacing between the peaks is more or less constant and fixed by the ratio of the comoving radius at decoupling to the comoving angular distance. The acoustic peaks therefore encode information on the cosmological parameters. In particular, for a Universe with $\Omega_\Lambda = 0$, $f_K(z_{\text{LSS}}) \sim 2/(H_0\Omega_0)$ and $r_s(z_{\text{LSS}}) \propto \Omega_0^{-1/2}$, so that for an adiabatic model, the position of the first peak gives almost directly an estimate of the curvature of the Universe,

$$\ell_{(1)} \sim \frac{220}{\sqrt{\Omega_0}}.$$

To finish, let us stress that the existence of these acoustic peaks was noted by Sakharov as early as 1965 [8], well before the observation of the temperature anisotropies. The acoustic peaks are therefore often called *Sakharov oscillations*.

6.2.3 Small scales

On small scales, k/τ' is no longer negligible. One should go beyond the tight coupling approximation as soon as $k \sim \tau'$, and the mean free path of photons becomes of the same order of magnitude as the wavelength of the perturbations.

6.2.3.1 Silk damping

Let us consider modes that have become sub-Hubble during the radiation era. Unlike the ones that come into play at intermediate scales, these modes have become sub-Hubble while the photons were dominating the matter content of the Universe. This implies that the gravitational potential has been greatly reduced during the oscillations. For a qualitative analysis, we can thus neglect the Bardeen potentials and forget about the gauge issue. Moreover, the characteristic time of acoustic oscillations is very small compared to the Hubble time so that one can neglect the effects of the expansion, i.e. assume R is constant.

The Euler equations (5.254) and (5.191) for the baryons and radiation then take the simplified form

$$V'_b = \frac{1}{R} \tau' (V_\gamma - V_b), \quad V'_\gamma = -\frac{1}{4} \delta_\gamma + \frac{1}{6} k^2 \pi_\gamma - \tau' (V_\gamma - V_b).$$

π_γ behaves as $1/\tau'$ and in order to go beyond the tight coupling limit we must determine the contribution in $1/\tau'$ of $(V_b - V_\gamma)$. Taking the difference of the two Euler equations the terms of order $(\tau')^0$ teach us that

$$\frac{R+1}{R} \tau' (V_\gamma - V_b) = -\frac{1}{4} \delta_\gamma. \quad (6.76)$$

Both Euler equations can then be combined to form $V'_\gamma + RV'_b = (R+1)V'_\gamma + R(V'_b - V'_\gamma)$. Using (6.76) to express $(V'_b - V'_\gamma)$, we obtain

$$(R+1)V'_\gamma = -\frac{1}{4}\delta_\gamma + \frac{1}{6}k^2\pi_\gamma - \frac{1}{4}\frac{R^2}{1+R}\frac{\delta'_\gamma}{\tau'}. \quad (6.77)$$

Including the effect of polarization, π_γ is related to V_γ by (6.58) [see (6.231)]. Combining this equation with the equation of continuity for the radiation, we then obtain

$$\delta''_\gamma + \frac{k^2 c_s^2}{\tau'} \left(\frac{16}{15} + \frac{R^2}{1+R} \right) \delta'_\gamma + k^2 c_s^2 \delta_\gamma = 0. \quad (6.78)$$

When neglecting the effect of polarization, the factor 16/15 should be replaced by a factor 4/5. The WKB solution of this equation is

$$\delta_\gamma \propto \exp \left(-\frac{k^2}{k_b^2} \right) \exp(\pm ikr_s), \quad (6.79)$$

with

$$k_b^{-2} = \frac{1}{6} \int_0^\eta \left[\frac{1}{1+R} \left(\frac{16}{15} + \frac{R^2}{1+R} \right) \right] \frac{d\eta'}{\tau'}. \quad (6.80)$$

The term in brackets varies very little during the cosmic history since it is 16/15 for $R=0$ and 1 for $R \rightarrow \infty$.

The mean free path for Compton scattering is $\lambda_C \sim 1/\tau'$. So the photons undergo a random walk that drifts in a time η of a length $\lambda_D \sim \sqrt{N}\lambda_C$, the number of steps being of the order of $N \sim \eta/\lambda_C$. The scattering length is thus of the order of $\lambda_D \sim \sqrt{\eta/\tau'}$. This is what (6.80) indicates: $\lambda_D \sim k_b^{-1} \sim \sqrt{\eta/\tau'}$. In a flat Universe with no cosmological constant, assuming that $z_{eq} \sim z_{LSS}$, it is of the order of

$$k_b \eta_{LSS} \sim (1+z_{LSS}) \sqrt{6 \frac{\sigma_T}{m_p} \frac{3H_0}{8\pi G_N} \Omega_{b0} h}. \quad (6.81)$$

This damping [9] leads to an exponential decrease of the angular power spectrum that starts to appear at around $\ell \sim 140\sqrt{\Omega_{b0}h^2}/0.019h$, i.e. before the first Doppler peak. The multipoles beyond $\ell \sim 2000$ are almost completely damped, which fixes the angular resolution that can be reached by observations. This effect is illustrated in Fig. 6.5.

6.2.3.2 Effect of the thickness of the last-scattering surface

As has been detailed in Chapter 4, recombination is not instantaneous and the surface of last scattering has a given thickness. The visibility function $g(\eta)$ or $g(z)$, characterizes the probability that a photon underwent its last scattering at a time η , or at

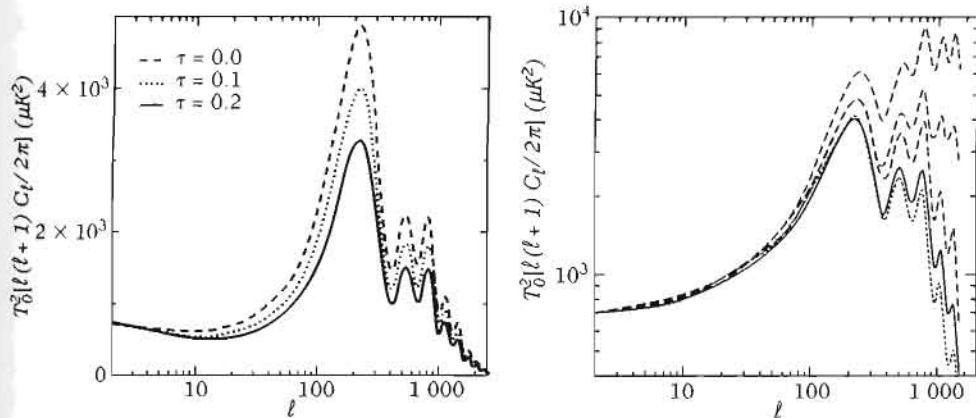


Fig. 6.5 (left): Effect of the reionization on the temperature angular power spectrum. (right): The dashes represent the anisotropies of a standard cold dark matter model without taking into account the Silk damping and the thickness of the surface of last-scattering surface (top) or the thickness of the surface of last scattering alone (middle) or the Silk effect alone (bottom). The solid line represents the exact computation and the dotted line corresponds to a computation including the effect of the thickness of the last scattering surface and the Silk effect by an analytical correction deduced from our estimates. From Ref. [3].

redshift z . The visibility function can be deduced from the optical depth of an object that emits its light at η , itself related to the differential opacity τ' by

$$\tau(\eta) = \int_{\eta}^{\eta_0} \tau'(u) du. \quad (6.82)$$

Notice the choice of sign in this expression that implies that $d\tau/d\eta = -\tau'!$ The visibility function is then

$$g(\eta) \equiv \tau' \exp(-\tau) \quad (6.83)$$

the integral of which is 1 by construction.

The temperature observed in a direction e is then an average weighted by the visibility function, of the form

$$\Theta(x_0, \eta_0, e) = \int g(\eta) [\Theta_{\text{SW}} + \Theta_{\text{dop}}](\eta, [\eta_0 - \eta]e) d\eta. \quad (6.84)$$

Moreover, the purely gravitational interactions to which the photons are subjected, at a given redshift, only affect the fraction of photons that underwent their last scattering. This fraction is expressed in terms of the visibility function as

$$g_\gamma(\eta) = \int_{\eta_{\text{in}}}^{\eta} g(u) du, \quad (6.85)$$

which varies from 0 to 1 between η_{in} and η_0 . The integrated Sachs-Wolfe term is then of the form

$$\Theta_{\text{ISW}} = \int g_\gamma(\eta) [\Phi' + \Psi'] (\eta, [\eta_0 - \eta] e) d\eta. \quad (6.86)$$

Assuming for simplicity that the temperature fluctuations are time independent then the expression (6.84) is simply the convolution, in the radial direction, of the temperature fluctuation with the visibility function. In Fourier space,

$$\Theta(x_0, \eta_0, e) \sim \int \hat{g}(k) (\hat{\Theta}_{\text{SW}} + \hat{\Theta}_{\text{dop}} + \hat{\Theta}_{\text{ISW}})(\eta_{\text{lss}}, k) e^{ik \cdot e \eta_0} \frac{dk}{\sqrt{2\pi}}.$$

As can be seen in Fig. 4.12, the visibility function can be described by a Gaussian function centred around η_{lss} ,

$$g(\eta) \sim \frac{1}{\sqrt{2\pi\sigma_{\text{lss}}^2}} \exp \left[-\frac{1}{2} \frac{(\eta - \eta_{\text{lss}})^2}{\sigma_{\text{lss}}^2} \right],$$

so that

$$\hat{g}(k) \propto \exp \left(-\frac{1}{2} k^2 \sigma_{\text{lss}}^2 \right). \quad (6.87)$$

The thickness of the last-scattering surface has the effect of exponentially washing out the temperature anisotropies on scales smaller than the thickness of the last-scattering surface. This effect is of the same order of magnitude as the Silk damping.

6.2.4 Other effects

After decoupling, the photons propagate almost freely so that only the gravitational effect can significantly change their energy. At lower redshifts, the Universe can be reionized, and the interaction with the plasma can also modify the photon temperature.

6.2.4.1 Integrated Sachs-Wolfe

The variation of the Bardeen potentials and primordial gravitational waves give a contribution to the temperature fluctuation, Θ_{ISW} , whose expression is given by (6.17). This component possesses two contributions:

- *Early Sachs-Wolfe effect*: the variation of the Bardeen potentials at the matter-radiation transition leaves a small contribution to the integrated term. For adiabatic models, this effect is dominant for the modes with wavelength entering the sound radius between equality and decoupling. Its contribution is thus dominant for multipoles close to the first peak. For these modes, $\Theta_{\text{SW}} \sim -\Phi/2$, which is of the same sign as $\Phi' < 0$ at the matter-radiation transition. Thus, Θ_{ISW} adds constructively to Θ_{SW} , hence inducing an increase in the amplitude of the first peak (Fig. 6.2). For isocurvature models, the growth of the potential for the modes with wavelength larger than the sound radius gives an important contribution for the multipoles smaller than the first peak.
- *The late Sachs-Wolfe effect* is due to the variation of the potential at small redshifts if the Universe is dominated by a cosmological constant or by the spatial curvature. This contribution is thus important mostly on large angular scales and it leads to a modification of the Sachs-Wolfe plateau since the potential is already

close to zero for the sub-Hubble modes at the time of decoupling. Since Φ decays (see Chapter 5), this induces an additional redshift. This effect is one of the rare signatures of dark energy on the anisotropies of the cosmic microwave background and it must be correlated with other tracers of the gravitational potential in the local Universe.

Another purely gravitational effect that can modify the angular power spectrum at small scales is that of gravitational lensing by large-scale structures. This effect is small on the temperature anisotropies but can have consequences for the polarization of the cosmic microwave background (see Chapter 7).

6.2.4.2 Reionization

As soon as the first sources of ultraviolet photons appear, the Universe can be reionized. The free electrons can then (re)scatter the photons of the cosmic microwave background after decoupling. This scattering tends to isotropize the anisotropies of the cosmic microwave background by averaging the temperature anisotropies from many lines of sight that meet at the scattering event. If we assume that reionization is homogeneous, we can approximate the visibility function by a sum of two Dirac distributions localized at η_{LSS} and η_{re} , respectively, for the decoupling and the reionization, then

$$\Theta(x_0, \eta_0, e) \simeq (1 - e^{-\tau_{\text{re}}}) (\Theta + e \cdot V_b) [e(\eta_0 - \eta_{\text{re}}), \eta_{\text{re}}] + e^{-\tau_{\text{re}}} (\Theta + e \cdot V_b) [e(\eta_0 - \eta_{\text{LSS}}), \eta_{\text{LSS}}], \quad (6.88)$$

where τ_{re} is the optical depth at reionization and where we have neglected the integrated Sachs-Wolfe effect.

The first term describes the averaging from different lines of sight as well as the generation of new anisotropies due to the scattering on moving electrons at reionization. The second term represents the temperature that would be observed if there were no reionization and has been weighted by the fraction of photons that have not been scattered.

$\Theta[e(\eta_0 - \eta_{\text{re}}), \eta_{\text{re}}]$, which is evaluated at the time of reionization, is the average of $\Theta - e' \cdot V_b$ on the electron last-scattering surface. For the long-wavelength modes, i.e. $k(\eta_{\text{re}} - \eta_0) \ll 1$, it gives $\Theta[e(\eta_0 - \eta_{\text{LSS}}), \eta_{\text{LSS}}]$, and it vanishes at small scales. Thus, for the super-Hubble scales at the time of reionization, the observed temperature anisotropy becomes

$$\Theta(x_0, \eta_0, e) \rightarrow \Theta(x_0, \eta_0, e) + (1 - e^{-\tau_{\text{re}}}) e \cdot \Delta V_b,$$

where ΔV_b is the difference between the electron velocities at reionization and at the previous scattering along the line of sight. At these scales, the Doppler term is subdominant so that $\Theta(e)$ is hardly affected. For sub-Hubble scales at the time of reionization

$$\Theta(x_0, \eta_0, e) \rightarrow e^{-\tau_{\text{re}}} \Theta(x_0, \eta_0, e) + (1 - e^{-\tau_{\text{re}}}) e \cdot V_b(\eta_{\text{re}}).$$

This effect is illustrated in Fig. 6.5. The analysis of the cross-correlation between temperature and polarization of the WMAP observations suggests that $\tau_{\text{re}} \sim 0.17$ from a redshift of $z_{\text{re}} \sim 15$.

6.2.4.3 Sunyaev–Zel'dovich effect

The interaction of the photons of the cosmic microwave background with a hot plasma can lead to a modification of their temperature. In particular, the hot gas of clusters has a temperature of the order of $10^7 - 10^8$ K, which radiates, via Bremsstrahlung, in the X-ray region. The inverse Compton scattering with the free electrons in the cluster can modify the energy distribution of photons by shifting the low-frequency photons to higher frequencies. This is the *thermal Sunyaev–Zel'dovich effect*, which represents the dominant contribution to the secondary anisotropies.

In a nutshell, the modification of the temperature of the cosmic microwave background photons is of the order of $\Delta T_{\text{SZ}} \sim \bar{L}(n_e T_e)$, where the bar stands for an average along the line of sight. L is the characteristic length of this line of sight in the cluster. T_e and n_e are the temperature and the density of the free electrons. Notice that the cluster surface brightness in X is of the order of

$$S_X \sim \frac{V}{4\pi D_L^2} \overline{(n_e n_p T_c^{1/2})},$$

where $V = D_A^2 \theta^2 L$ is the volume of the cluster, assumed to be observed under an angular diameter θ .

To illustrate this effect, let us consider a cluster with electronic density profile of the form

$$n_e(r) = n_0 \left[1 + \left(\frac{r}{r_c} \right)^2 \right]^{-3\beta/2}, \quad (6.89)$$

for $0 < r < R_{\text{cluster}}$ and zero otherwise, where r_c is the core radius. The variation of the temperature of the photon is given by

$$\Delta T_{\text{SZ}} = -2 \frac{k_B T_0}{m_e c^2} \sigma_T \int_{-L_{\max}}^{L_{\max}} n_e dl, \quad (6.90)$$

assuming that the temperature of the hot gas is independent of r . In the limit $R_{\text{cluster}} \rightarrow \infty$, we infer that

$$\Delta T_{\text{SZ}} = -2 \frac{k_B T_0}{m_e c^2} \sigma_T n_0 r_c B \left(\frac{3\beta - 1}{2}, \frac{1}{2} \right) \left[1 + \left(\frac{\theta}{\theta_c} \right)^2 \right]^{(1-\epsilon\beta)/2}. \quad (6.91)$$

Unlike Thomson scattering, the thermal Sunyaev–Zel'dovich effect induces a spectral distortion with a temperature decrease at long wavelengths and an increase at short ones. It allows us to probe galaxy clusters and represents a powerful tool to detect clusters at large redshifts (see Ref. [10] for a review). If the cluster has velocity v with respect to the cosmic microwave background, the Doppler effect induces a second effect, called the *kinetic Sunyaev–Zel'dovich effect*, which depends on the radial component of the velocity.

6.2.4.4 Gravitational waves

It is difficult to obtain a relation equivalent to (6.50) for gravitational waves since (6.45) involves the time derivative of \bar{E}_λ . In the matter era, the solution of the gravitational wave equation is given by (5.135) so that

$$\bar{E}'_\lambda(k, \eta) = -3k\bar{E}_\lambda(k, \eta_i)\frac{j_2(k\eta)}{k\eta}. \quad (6.92)$$

The relation (6.45) then implies that, for a primordial spectrum of the form $\mathcal{P}_T \propto k^{n_T}$ as predicted by inflation (8.234),

$$C_\ell^T \propto \frac{(\ell+2)!}{(\ell-2)!} \int \frac{dy_0}{y_0^{1-n_T}} \left[\int_{y_0\eta_{\text{LSS}}/\eta_0}^{y_0} \frac{j_\ell(y_0-y) j_2(y)}{(y_0-y)^2} dy \right]^2, \quad (6.93)$$

where we have set $y = k\eta$ and chosen $k_0 = 1/\eta_0$. This integral is non-negligible only when $y \sim 2$ and $y_0 - y \sim \ell$ simultaneously, since otherwise at least one of the two Bessel functions is close to zero. The integration domain implies that $y_0 < y\eta_0/\eta_{\text{LSS}}$, so that both conditions can be satisfied simultaneously only if

$$\ell \lesssim 2 \frac{\eta_0}{\eta_{\text{LSS}}} \sim 2\sqrt{z_{\text{LSS}} + 1} \sim 60. \quad (6.94)$$

Figure 6.3 indicates that C_ℓ^T decreases rapidly from $\ell \sim 100$ onwards, which is compatible with this estimate. This implies that the spectrum C_ℓ^T is weakly sensitive to the cosmological parameters.

To evaluate the integral (6.93), notice that the second Bessel function is peaked around $y \sim 2$. It can be approximated by $j_2(y)/y \sim 2^{-1/2}\delta(y-2)$, which implies, after integrating over y , that

$$C_\ell^T \propto \frac{(\ell+2)!}{(\ell-2)!} \int \frac{dy_0}{y_0^{1-n_T}} \left[\frac{j_\ell(y_0)}{y_0^2} \right]^2, \quad (6.95)$$

when we assume that $y_0 \gg 2$. We infer that

$$C_\ell^T \propto \frac{(\ell+2)!}{(\ell-2)!} \frac{\Gamma\left(\ell-2+\frac{n_T}{2}\right)}{\Gamma\left(\ell+4+\frac{n_T}{2}\right)}. \quad (6.96)$$

Using the same expansion as for (6.50), we obtain

$$\ell(\ell+1)C_\ell^T \propto \ell^{n_T} \quad (6.97)$$

for small multipoles, $\ell \gtrsim 2$.

6.3 Kinetic description

To obtain a more precise and realistic description of the photon propagation and of their decoupling from matter, we must develop a kinetic approach. This approach is mainly necessary for relativistic fluids, which have a large mean free path. The kinetic theory in the relativistic framework is described in detail in Refs. [11–13] and complementary material on distribution functions can be found in Ref. [14]. The Boltzmann equation in the cosmological framework is detailed in Refs. [15–19].

6.3.1 Perturbed Boltzmann equation

6.3.1.1 Distribution function

Just as in Chapter 5, we introduce the distribution function $f(x^\mu, p_\mu)$ that depends on the position and the momentum, which are considered to be independent. For a particle of mass m , the distribution function lives on the mass shell

$$\mathcal{P}_m = \{(x^\mu, p_\mu) \in T\mathcal{M}, \quad p^\mu p_\mu = -m^2\}, \quad (6.98)$$

where the tangent space is defined as

$$T\mathcal{M} = \{(x^\mu, p_\mu), x^\mu \in \mathcal{M}, p_\mu \in T_x\},$$

\mathcal{M} being the four-dimensional manifold describing space-time and T_x its tangent space at the point x .

To decompose this distribution function, it is convenient to introduce the vector field normal to the constant time hypersurfaces, $\{\eta = \text{const.}\}$. With a metric of the form (5.52), this vector field can be expressed as

$$N_\mu = -a(1 + A, 0), \quad N^\mu = \frac{1}{a}(1 - A, -B^i), \quad (6.99)$$

to first order in perturbations. One can then decompose the tangent vector k^μ to the geodesics of the photons (we shall use p^μ for the general notation and k^μ only for null vectors) as

$$k^\mu = \frac{E}{a} \left[1 - A, n^i - \left(B_j n^j + \frac{1}{2} h_{jk} n^j n^k \right) n^i \right], \quad (6.100)$$

with $\gamma_{ij} n^i n^j = 1$. This decomposition ensures firstly that $k^\mu k_\mu = 0$ and secondly it allows us to identify E with the energy of the photon that would be measured by an observer comoving with the four-velocity N^μ since $N_\mu k^\mu = -E$.

In perturbation theory, the distribution function can then be decomposed into a part that, because of isotropy and homogeneity, depends only on η and E and into a perturbation, that will depend on the space-time position and the 4 components of the momentum, as

$$f(x^\mu, p_\mu) = f(\eta, x^i, E, n^j) = \bar{f}(\eta, E) + \delta f(\eta, x^i, E, n^j). \quad (6.101)$$

The evolution of this distribution function is dictated by the Boltzmann equation

$$L[f] = C[f], \quad (6.102)$$

where $L \equiv d/d\eta$ is the Liouville operator and $C[f]$ the collision term. This equation implies that the photon density in phase space is only affected by collisions.

6.3.1.2 Collisionless Boltzmann equation

To start with, let us consider the Liouville term,

$$L[f] = \frac{df}{d\eta} = \left(k^\mu \frac{\partial f}{\partial x^\mu} + \frac{dk^\mu}{d\lambda} \frac{\partial f}{\partial k^\mu} \right) \frac{d\lambda}{d\eta}, \quad (6.103)$$

where λ is an affine parameter along the geodesic with tangent vector k^μ , so that $d\eta/d\lambda = k^0$. The geodesic equation (1.37) can be rewritten as

$$\frac{dk^\mu}{d\lambda} = -\Gamma_{\alpha\beta}^\mu k^\alpha k^\beta, \quad (6.104)$$

so that the Liouville operator takes the general form

$$L[f] = \left(k^\mu \frac{\partial f}{\partial x^\mu} - \Gamma_{\alpha\beta}^\mu k^\alpha k^\beta \frac{\partial f}{\partial k^\mu} \right) \frac{1}{k^0}. \quad (6.105)$$

From the decomposition (6.101) it can be rewritten as

$$L[f] = \frac{\partial f}{\partial \eta} + \frac{dx^i}{d\eta} \frac{\partial f}{\partial x^i} + \frac{dE}{d\eta} \frac{\partial f}{\partial E} + \frac{dn^i}{d\eta} \frac{\partial f}{\partial n^i}. \quad (6.106)$$

Now, we must evaluate this function to first order in the perturbations. The last term is easy to obtain since $\partial f / \partial n^i$ is of order 1 in perturbations so that $dn^i/d\eta$ should only be evaluated at the background level. To this order, (6.104) implies that

$$E' = -\mathcal{H}E, \quad n^{i'} = -(^{(3)}\Gamma_{jk}^i n^j n^k), \quad (6.107)$$

$^{(3)}\Gamma_{jk}^i$ being the Christoffel symbols of the spatial metric γ_{ij} (see Appendix A). The last term of (6.106) is thus

$$\frac{dn^i}{d\eta} \frac{\partial f}{\partial n^i} = -(^{(3)}\Gamma_{jk}^i n^j n^k) \frac{\partial f}{\partial n^i}.$$

The second term is also very simple to obtain since $dx^i/d\eta = k^i/k^0$. Since $\partial_i f$ is of first order in perturbations, we only need $dx^i/d\eta = n^i$ so that

$$\frac{dx^i}{d\eta} \frac{\partial f}{\partial x^i} = n^i \partial_i f.$$

To evaluate the third term of the expression (6.106), one should compute $dE/d\eta$. For this it is sufficient to consider the time component of (6.104),

$$\frac{dk^0}{d\eta} = -\Gamma_{\alpha\beta}^0 \frac{k^\alpha k^\beta}{k^0},$$

which directly gives

$$\frac{d}{d\eta} \left(\frac{E}{a} \right) = \frac{E}{a} (A' + n^i \partial_i A) - \Gamma_{\alpha\beta}^0 \frac{k^\alpha k^\beta}{E/a} (1 + 2A).$$

Using the expressions (C.6) and (C.26) for the Christoffel symbols, we obtain

$$\frac{dE}{d\eta} = -E \left[\mathcal{H} + n^i \partial_i A + \left(\frac{1}{2} h'_{ij} - D_{(i} B_{j)} \right) n^i n^j \right].$$

Gathering all these terms together, we then obtain the final form of the Liouville operator to first order in the perturbations

$$L[f] = f' + n^k \partial_k f - \left[\mathcal{H} + n^k \partial_k A + \left(\frac{1}{2} h'_{ij} - D_{(i} B_{j)} \right) n^i n^j \right] E \frac{\partial f}{\partial E} - {}^{(3)}\Gamma_{ij}^k n^i n^j \frac{\partial f}{\partial n^k}. \quad (6.108)$$

$$L[f] = f' + n^k \partial_k f - \left[\mathcal{H} + n^k \partial_k A + \left(\frac{1}{2} h'_{ij} - D_{(i} B_{j)} \right) n^i n^j \right] E \frac{\partial f}{\partial E} - {}^{(3)}\Gamma_{ij}^k n^i n^j \frac{\partial f}{\partial n^k}. \quad (6.109)$$

With the decomposition (6.101), we thus obtain the two evolution equations

$$L[\bar{f}] = \bar{f}' - \mathcal{H} E \frac{\partial \bar{f}}{\partial E} \quad (6.110)$$

$$L[\delta f] = \delta f' + n^k \partial_k \delta f - \mathcal{H} E \frac{\partial \delta f}{\partial E} - \left[n^k \partial_k A + \left(\frac{1}{2} h'_{ij} - D_{(i} B_{j)} \right) n^i n^j \right] E \frac{\partial \bar{f}}{\partial E} - {}^{(3)}\Gamma_{ij}^k n^i n^j \frac{\partial \delta f}{\partial n^k}. \quad (6.111)$$

To lowest order, the distribution function of the photons is a Bose-Einstein distribution function (Chapter 4),

$$\bar{f}(E, \eta) = \left\{ \exp \left[\frac{E}{T(\eta)} \right] - 1 \right\}^{-1} = \bar{f} \left(\frac{E}{T} \right). \quad (6.112)$$

(6.110) then implies that $T \propto 1/a$, as was already given for the energy of each photon by the geodesic equation.

To linear order, if we consider the distribution function to be of the form

$$f = \bar{f} \left[\frac{E}{T(\eta, x^i, n^j)} \right],$$

and expand the temperature as $T(\eta, x^i, n^j) = \bar{T}(\eta)[1 + \theta(\eta, x^i, n^j)]$ then

$$f(\eta, x^i, E, n^j) = \bar{f}(\eta, E) - E \frac{\partial \bar{f}}{\partial E} \theta(\eta, x^i, n^j). \quad (6.113)$$

The temperature field is thus inhomogeneous (it depends on x^i) and anisotropic (it depends on n^j). We stress that this ansatz implicitly assumes that θ does not depend

on E . This hypothesis can only be justified after the collision term has been studied. As we shall see, this term does not depend on the energy, so the Thomson scattering, in this regime, does not induce any spectral distortion.

6.3.1.3 Macroscopic quantities

A series of macroscopic quantities can be defined from the distribution function, the first of which are fluid quantities. For this we introduce a space-time tetrad, or vierbein, $(e_\mu^a)_{a=0..3}$, to be locally in a Minkowski space-time. We choose this tetrad such that

$$e_\mu^0 = N_\mu, \quad e_\mu^i e_\nu^j g^{\mu\nu} = \delta^{ij}, \quad e_\mu^i N^\mu = 0.$$

The vector k^μ can then be decomposed as $k^a = k^\mu e_\mu^a$. The energy-momentum tensor is then obtained, as in Minkowski space-time, by integrating over the momentum as

$$T^{\mu\nu}(x) = \int_{\mathcal{P}_m(x)} k^\mu k^\nu f(x^\alpha, k^\beta) \frac{d^3k}{(2\pi)^3 E_k}, \quad (6.114)$$

with $k^2 = \delta_{ab} k^a k^b$ and $E_k = \sqrt{k^2 + m^2} = k$ for a massless particle. To simplify the notations, the factor $(2\pi)^3$ is included in the distribution function. The energy-momentum tensor of radiation is then given by

$$T^{\mu\nu}(x) = \int_{\mathcal{P}_m(x)} k^\mu k^\nu f(x^\alpha, k^\beta) E dE d^2n. \quad (6.115)$$

This energy-momentum tensor can be expanded as (5.69). We then infer that

$$\rho = \int (k^\mu u_\mu)^2 f E dE d^2n, \quad (6.116)$$

$$P = \frac{1}{3} \int (k^\mu k^\nu \perp_{\mu\nu}) f E dE d^2n, \quad (6.117)$$

$$\Pi^{\mu\nu} = \int k^\alpha k^\beta \left(\perp_\alpha^\mu \perp_\beta^\nu - \frac{1}{3} \perp_{\alpha\beta} \perp^{\mu\nu} \right) f E dE d^2n, \quad (6.118)$$

where we recall that $\perp_{\mu\nu} = g_{\mu\nu} + u_\mu u_\nu$. u^μ is defined by (5.70) and thus $u_\mu = a(-1 - A, v_k + B_k)$. $k^\mu k_\mu = 0$ implies that $P = \rho/3$ and we recover the equation of state for radiation. We first note that

$$k^\mu u_\mu = -E [1 - n^i(v_i + B_i)],$$

and then notice that the integral of $\bar{f} n^i$ over angles vanishes, so that we infer that

$$\delta\rho = \int \delta f E^3 dE d^2n, \quad (6.119)$$

$$(\rho + P)(v^i + B^i) = \int n^i \delta f E^3 dE d^2n, \quad (6.120)$$

$$\Pi^{ij} = P \pi^{ij} = \int n^{ij} \delta f E^3 dE d^2n, \quad (6.121)$$

with

$$n_{ij} \equiv n_i n_j - \frac{1}{3} \gamma_{ij}. \quad (6.122)$$

6.3.1.4 Brightness and temperature

The brightness, I , is defined by integrating the distribution function over the energy

$$I(\eta, x^i, n^j) \equiv 4\pi \int E^3 f(\eta, x^i, E, n^j) dE. \quad (6.123)$$

It therefore represents the energy density per unit solid angle that propagates along a direction n^i at the space-time point (η, x^j) . This quantity can be decomposed as $I = \bar{I} + \delta I$ with

$$\bar{I}(\eta) \equiv 4\pi \int E^3 \bar{f}(\eta, E) dE, \quad \delta I(\eta, x^i, n^j) \equiv 4\pi \int E^3 \delta f(\eta, x^i, E, n^j) dE. \quad (6.124)$$

Notice that from (6.116), the photon energy density, ρ_γ , is simply the monopole of this brightness, i.e.

$$\rho_\gamma(\eta, x^j) = \int \frac{d^2 n}{4\pi} I(\eta, x^i, n^j), \quad (6.125)$$

and (6.110) then implies that

$$\bar{I}' + 4\mathcal{H}\bar{I} = 0, \quad \bar{\rho}'_\gamma + 4\mathcal{H}\bar{\rho}_\gamma = 0 \quad (6.126)$$

and also that $\bar{I} = \bar{\rho}_\gamma$, which is obvious since \bar{I} does not depend on n^i because of the background space-time symmetries.

Similarly, for the collision term entering (6.102), we define

$$C[\delta J] = 4\pi \int C[\delta f] E^3 dE. \quad (6.127)$$

Integrating the (6.111) over the energy, we then obtain

$$\begin{aligned} \delta I' + n^k \partial_k \delta I + 4\mathcal{H}\delta I + 4 \left[n^k \partial_k A + \left(\frac{1}{2} h'_{ij} - D_{(i} B_{j)} \right) n^i n^j \right] \bar{I} \\ - {}^{(3)}\Gamma_{ij}^k n^i n^j \frac{\partial \delta J}{\partial n^k} = C[\delta J]. \end{aligned} \quad (6.128)$$

The temperature contrast is defined from the brightness by

$$\theta(\eta, x^i, n^i) = \frac{1}{4} \frac{\delta I}{\bar{I}}, \quad (6.129)$$

and corresponds to the definition introduced by the (6.113). The monopole of this quantity is directly related to the contrast of the temperature of the photons

$$\theta_0(\eta, x^i) = \int \theta(\eta, x^i, n^i) \frac{d^2 n}{4\pi} = \frac{1}{4} \delta_\gamma. \quad (6.130)$$

The evolution equation for θ can be easily inferred from that of the brightness (6.128), using (6.126),

$$\theta' + n^k \partial_k \theta - {}^{(3)}\Gamma_{ij}^k n^i n^j \frac{\partial \theta}{\partial n^k} + \left[n^k \partial_k A + \left(\frac{1}{2} h'_{ij} - D_{(i} B_{j)} \right) n^i n^j \right] = C[\theta], \quad (6.131)$$

where we have defined $C[\theta] = C[\delta I]/4\bar{I}$. This equation can be obtained directly by inserting the decomposition (6.113) in (6.111). However, a difference exists since here θ represents the relative fluctuation of the brightness and the integration over the energy implies that this quantity does not depend on the temperature, an hypothesis implicit in the definition (6.113). Both quantities are identical and satisfy the same Boltzmann equation as long as the collision term does not depend on the energy.

6.3.1.5 Multipole expansion

Just as in (6.125), the different fluid quantities appear as the successive moments of brightness

$$(\rho + P)(v^i + B^i) = \int \delta I n^i \frac{d^2 n}{4\pi}, \quad \Pi^{ij} = \int n^{ij} \delta I \frac{d^2 n}{4\pi}. \quad (6.132)$$

We can thus expand the Boltzmann (6.128) by expanding the brightness in a series as

$$\delta I(\eta, x^i, n^j) = \sum_{\ell} (n^{i_1} \dots n^{i_\ell}) \mathcal{I}_{i_1 \dots i_\ell}(\eta, x^i). \quad (6.133)$$

The only metric perturbations that are not gauge invariant are scalar and vector modes so that only \mathcal{I}_0 and \mathcal{I}_{i_1} are modified under a gauge transformation. All higher-order components are gauge invariant. Such a decomposition is not unique but this problem can be fixed by imposing the quantities $\mathcal{I}_{i_1 \dots i_\ell}$ to be symmetric and traceless. Under these conditions $\mathcal{I}_{i_1 \dots i_\ell}$ possesses only $(2\ell + 1)$ independent components [only $(\ell + 1)(\ell + 2)/2$ out of the 3^ℓ components of $\mathcal{I}_{i_1 \dots i_\ell}$ are independent if the tensor is symmetric and the condition on the trace imposes $\ell(\ell - 1)/2$ constraints among these components]. As seen previously, only the first three terms of this expansion are fluid quantities, and as such will enter the Einstein equations. A similar series expansion can be performed for the temperature field, leading to the definition of the coefficients $\theta_{i_1 \dots i_\ell}(\eta, x^i)$.

Notice that the expansion (6.133) allows us to simplify the formulation of the Boltzmann equation since in $n^k \partial_k \delta I$, the derivative is only applied on $\mathcal{I}_{i_1 \dots i_\ell}$, so that this term can be combined with the Christoffel symbols to lead to a (spatial) covariant derivative. Besides, this is the only proper way to deal with terms such as $\partial \theta / \partial n^k$. Equation (6.131) then takes the form

$$\theta' + \sum_{\ell} n^k (n^{i_1} \dots n^{i_\ell}) D_k \theta_{i_1 \dots i_\ell} + \left[n^k \partial_k A + \left(\frac{1}{2} h'_{ij} - D_{(i} B_{j)} \right) n^i n^j \right] = C[\theta]. \quad (6.134)$$

The metric terms do not depend on n^i so that they only come into play in the equations for the monopole and the dipole, i.e. at the level of the fluid description.

The SVT decomposition, introduced in Chapter 5, can be generalized to all orders. We recall that a trace-free symmetric tensor of rank 2, such as the anisotropic stress tensor (5.72), can be decomposed as

$$\pi_{ij} = \Delta_{ij}\pi + D_{(i}\bar{\pi}_{j)} + \bar{\pi}_{ij},$$

where we recall that Δ_{ij} is defined by (5.73). We thus obtain the scalar, vector and tensor parts as

$$\pi_{ij}^{(S)} = [D_{ij}\pi]^\text{STF}, \quad \pi_{ij}^{(V)} = [D_i\pi_j]^\text{STF}, \quad \pi_{ij}^{(T)} = [\pi_{ij}]^\text{STF},$$

where ‘STF’ means that we extract the symmetric and trace-free part. For the decomposition to be unique, it is clear that we should impose $\pi_{ij}^{(V)}$ and $\pi_{ij}^{(T)}$ to be traceless and transverse.

This decomposition can be generalized to any tensor of rank ℓ and the component of order (m) can be decomposed as

$$\hat{\theta}_{i_1\dots i_\ell}^{(m)} = \left[D_{i_1\dots i_{\ell-m}} \hat{\theta}_{i_{\ell-m+1}\dots i_\ell}^{(m)} \right]^\text{STF}. \quad (6.135)$$

$|m| = 0, 1, 2$ correspond, respectively, to the terms S, V and T. Decomposing θ in Fourier coefficients, $\hat{\theta}$, and choosing a basis such that $e_3 = k/k$, then the scalar part of $\hat{\theta}_{i_1\dots i_\ell}$ will be proportional to $k^{i_1}\dots k^{i_\ell}\hat{\theta}_\ell^{(0)}$. Contracting it with the SFT part of $n^{i_1}\dots n^{i_\ell}$, it is clear that we obtain a Legendre polynomial of order ℓ so that

$$n^{i_1}\dots n^{i_\ell}\hat{\theta}_{i_1\dots i_\ell}^{(0)} \propto Y_{\ell 0}\hat{\theta}_\ell^{(0)}.$$

The vector modes have two polarizations that are vectors orthogonal to k [see the decomposition (6.32)] so we can choose $e^\pm = e_1 \pm e_2$ as the basis for the vector modes. The vector part of $\hat{\theta}_{i_1\dots i_\ell}$ is then proportional to the STF part of $k_{i_1}\dots k_{i_{\ell-1}}e_{i_\ell}^\pm\hat{\theta}_\ell^{(\pm 1)}$. In this case, the contraction with $n^{i_1}\dots n^{i_\ell}$ gives terms of the form

$$n^{i_1}\dots n^{i_\ell}\hat{\theta}_{i_1\dots i_\ell}^{(\pm 1)} \propto Y_{\ell \pm 1}\hat{\theta}_\ell^{(\pm 1)}.$$

This construction can be generalized for all m . Thus, the Fourier component $\hat{\theta}_{i_1\dots i_\ell}$ can be decomposed as

$$\hat{\theta}_{i_1\dots i_\ell} = \sum_{m=-\ell}^{\ell} \hat{\theta}_{i_1\dots i_\ell}^{(m)} \quad (6.136)$$

with

$$\hat{\theta}_{i_1\dots i_\ell}^{(m)} \propto \left[k_{i_1}\dots k_{i_{\ell-m}} e_{i_{\ell-m+1}}^\pm \otimes \dots \otimes e_{i_\ell}^\pm \right]^\text{STF} \hat{\theta}_\ell^{(m)}, \quad (6.137)$$

choosing the + sign when $m \geq 0$ and the - sign otherwise. One can then show (see Chapter 5 of Ref. [3] or Ref. [20], which contains a description of the construction of the STF tensors) that

$$n^{i_1}\dots n^{i_\ell}\hat{\theta}_{i_1\dots i_\ell}^{(m)} \propto Y_{\ell m}\hat{\theta}_\ell^{(m)}. \quad (6.138)$$

Each type of perturbation possesses 2 degrees of freedom apart from the scalar mode that only has one, so that for a mode ℓ , there are $2\ell + 1$ degrees of freedom.

This construction allows us to conclude that the temperature field $\theta(\eta, \mathbf{x}, \mathbf{n})$ of radiation at the point \mathbf{x} that propagates in the direction \mathbf{n} can be expanded in the basis $G_{\ell m}(\mathbf{x}, \mathbf{n})$ defined by

$$G_{\ell m}(\mathbf{x}, \mathbf{n}) = (-i)^\ell \sqrt{\frac{4\pi}{2\ell+1}} e^{i\mathbf{k}\cdot\mathbf{x}} Y_{\ell m}(\mathbf{n}), \quad (6.139)$$

as

$$\theta(\eta, \mathbf{x}, \mathbf{n}) = \int \frac{d^3 k}{(2\pi)^{3/2}} \sum_{\ell m} \hat{\theta}_\ell^{(m)}(\eta, \mathbf{k}) G_{\ell m}(\mathbf{x}, \mathbf{n}). \quad (6.140)$$

Only the modes $m = 0, \pm 1, \pm 2$ are excited by the matter and space-time geometry perturbations.

6.3.1.6 Relation with the SVT decomposition

Expanding the scalar modes as

$$Q^{(0)} = \exp(i\mathbf{k}\cdot\mathbf{x}), \quad Q_i^{(0)} = \partial_i Q^{(0)}, \quad Q_{ij}^{(0)} = -\left(\partial_i \partial_j - \frac{1}{3} \delta_{ij} \Delta\right) Q^{(0)},$$

then

$$Q^{(0)} = G_{00}, \quad n^i Q_i^{(0)} = -k G_{10}, \quad n^i n^j Q_{ij}^{(0)} \propto k^2 G_{20}. \quad (6.141)$$

As for the vector modes, they can be expanded as

$$Q_i^{(\pm 1)} = -\frac{i}{\sqrt{2}} e_i^\pm Q^{(0)}, \quad Q_{ij}^{(\pm 1)} = \partial_{(i} Q_{j)}^{(\pm 1)},$$

so that

$$n^i Q_i^{(\pm 1)} = G_{1\pm 1}, \quad n^i n^j Q_{ij}^{(\pm 1)} \propto -k G_{2\pm 1}. \quad (6.142)$$

This decomposition is analogous to that of (6.32) which we have used initially. The tensor modes can be expanded as

$$Q_{ij}^{(\pm 2)} = -\sqrt{\frac{3}{8}} e_i^\pm \otimes e_j^\pm Q^{(0)},$$

so that

$$n^i n^j Q_{ij}^{(\pm 2)} = G_{2\pm 2}. \quad (6.143)$$

Notice that this expansion differs from that initially used (6.40) by a factor $\sqrt{3}/8$.

6.3.1.7 Collision term

In general, the collision term can be decomposed as

$$C[f] = \frac{df_+}{d\eta} - \frac{df_-}{d\eta} \quad (6.144)$$

where f_+ and f_- represent the incoming and outgoing distribution functions during the scattering. In the rest frame of the baryons and electrons, the two terms involved in (6.144) are given by

$$\frac{df_+}{d\eta} = \tau' \int f \omega(n, n') \frac{d^2 n'}{4\pi}, \quad \frac{df_-}{d\eta} = \tau' f, \quad (6.145)$$

where τ' is defined as previously as

$$\tau' = a\sigma_T n_e. \quad (6.146)$$

σ_T is the cross-section of the Thomson scattering and n_e the density of free electrons. The function $\omega(n, n')$ contains the angular dependence of the Thomson scattering and is given by

$$\omega(n, n') = \frac{3}{4} [1 + (n \cdot n')^2] = 1 + \frac{3}{4} n^{ij} n'_{ij} = 1 + \frac{1}{2} P_2(n \cdot n'), \quad (6.147)$$

P_2 being the Legendre polynomial of order 2. Notice an important property: ω and thus the collision term does not depend on the energy so that the Thomson scattering does not induce any spectral distortion. This is no longer the case when the temperature of the photons is very different from that of the electrons, as, for instance, for the Sunyaev-Zel'dovich effect clusters where the temperature of the electrons is of the order of a few keV, while that of the photons is only a few Kelvin. This property also implies that the brightness fluctuation (6.129) does indeed correspond to a temperature fluctuation, as introduced by the ansatz (6.113).

At the background level, the expression for $\omega(n, n')$ implies that

$$C[\bar{f}] = 0, \quad (6.148)$$

which is to be expected since at this order the symmetries of the Friedmann-Lemaître space-time impose that all fluids have the same velocity, proportional to δ_0^μ .

To first order in the perturbations, we get

$$C[\delta f] = \tau' \left[\int \delta f(\eta, x^i, \bar{E}, n') \frac{d^2 n'}{4\pi} - \delta f(\eta, x^i, \bar{E}, n) + \frac{3}{4} n^{ij} \int \delta f(\eta, x^i, \bar{E}, n') n'_{ij} \frac{d^2 n'}{4\pi} \right], \quad (6.149)$$

where \bar{E} is the energy in the rest frame of the electrons and the baryons. Using the fact that $\bar{E} = -k_\mu u_b^\mu = E[1 - (B^i + v_b^i)n_i]$, we obtain that

$$E^3 dE = [1 + 4(B^i + v_b^i)n_i] \bar{E}^3 d\bar{E}.$$

The integral over the energy then gives

$$C[\delta I] = \tau' \left[\int \delta I \frac{d^2 n'}{4\pi} - \delta I + 4\bar{I}(B^i + v_b^i)n_i + \frac{3}{4}n^{ij} \int \delta I(\eta, x^i, \bar{E}, n') n'_{ij} \frac{d^2 n'}{4\pi} \right]. \quad (6.150)$$

The first term reduces to the monopole and the last one to $3n^{ij}\Pi_{ij} = (4\bar{I})n^i n^j \pi_{ij}/16$, so that

$$C[\theta] = \tau' \left[\theta_0 - \theta + (B^i + v_b^i)n_i + \frac{1}{16}n^i n^j \pi_{ij} \right], \quad (6.151)$$

where θ_0 is the monopole of the temperature (6.130). Notice again that $C[\theta]$ vanishes in the unperturbed space-time. From the Stewart–Walker lemma (Chapter 5), $C[\theta]$ is thus gauge invariant. We also stress that in the expansion scheme (6.133), the only term that contributes beyond the multipole $\ell = 2$ is $C[\theta] = -\tau'\theta$. So, for $\ell > 2$ no coupling term to the baryons or to the metric will contribute to the Boltzmann hierarchy.

6.3.2 Gauge invariant expressions

Some complementary material on the discussion of the gauge dependence of the distribution function can be obtained in Refs. [19, 21].

6.3.2.1 Gauge invariant distribution function

The decomposition of the distribution function f as $f = \bar{f} + \delta f$ is not unique. Just as in Chapter 5, the definition of the perturbation δf implies the choice of an isomorphism ψ between the spaces where \bar{f} and f are defined, i.e. of a map

$$\psi : \mathcal{P}_m \mapsto \bar{\mathcal{P}}_m : (x^\alpha, p^\beta) \mapsto (\bar{x}^\alpha, \bar{p}^\beta),$$

which allows us to decompose f as

$$f = \bar{f} \circ \psi + \delta f. \quad (6.152)$$

Any change of coordinates, $x^\alpha \rightarrow y^\alpha = x^\alpha - \xi^\alpha$, on \mathcal{M} induces a transformation on the tangent space in which the momentum transforms as

$$p^\mu \longrightarrow p^\mu - p^\nu \partial_\nu \xi^\mu. \quad (6.153)$$

This transformation is thus generated by the vector with coordinates

$$T\xi = (\xi^\mu, p^\nu \partial_\nu \xi^\mu) \quad (6.154)$$

in the natural basis $(\partial_\mu, \partial_{p^\mu})$. To linear order, the distribution function transforms as

$$f \longrightarrow f + \mathcal{L}_{T\xi} f. \quad (6.155)$$

To obtain the transformation law of δf , one should first notice that \bar{f} is not defined on $T\mathcal{M}$ but only on the subspace $\bar{\mathcal{P}}_m$. But in general, $T\xi$ is not tangent to $\bar{\mathcal{P}}_m$, so that $\mathcal{L}_{T\xi} \bar{f}$ is not well defined. \bar{f} should therefore be extended on an open set of $T\mathcal{M}$

containing $\bar{\mathcal{P}}_m$. Performing a gauge transformation and keeping only the first order terms, we obtain

$$\bar{f}_1 \circ \psi_1 + \delta f_1 \rightarrow \bar{f}_1 \circ \psi_1 + \delta f_1 + \mathcal{L}_{T\xi} \bar{f}_1,$$

since $\mathcal{L}_{T\xi} \delta f_1$ is of second order. However, one should notice that there can exist several decompositions of the same function, $f = \bar{f}_1 \circ \psi_1 + \delta f_1 = \bar{f}_2 \circ \psi_2 + \delta f_2$. In this case, there exists a vector field ζ such that $\bar{f}_2 = \bar{f}_1 - \mathcal{L}_{T\xi} \bar{f}_1$ so that

$$\delta f_2 - \delta f_1 = \bar{f}_1 \circ \psi_1 - \bar{f}_1 \circ \psi_2 + \mathcal{L}_{T\xi} \bar{f}_1.$$

The variation of δf under a change of coordinates is thus given by

$$\bar{f}_1 \circ \psi_1 - \bar{f}_1 \circ \psi_2 + \mathcal{L}_{T\xi} \bar{f}_1 \equiv \mathcal{L}_{T\xi\parallel} \bar{f}.$$

One can show that $T\xi\parallel$ is parallel to $\bar{\mathcal{P}}_m$ so that $\mathcal{L}_{T\xi\parallel} \bar{f}$ does not depend on the extension of \bar{f} . The computation of this term (see Ref. [21]) allows us to show that under a gauge transformation, the perturbation of the distribution function transforms as²

$$\delta f \longrightarrow \delta f + (\mathcal{H}T + n^i \partial_i T) E \frac{\partial \bar{f}}{\partial E},$$

(6.156)

where we have used the notations of Chapter 5. A more intuitive way to obtain this result is to remind ourselves, (6.101), that δf is a function of (η, x^i, E, n^j) , where from (6.99), E is given by $E = -N_\mu k^\mu$ so that

$$E = a(1 + A)k^0.$$

Under a coordinate transformation of the form $x^\alpha \rightarrow x^\alpha - \xi^\alpha$, k^μ transforms as (6.153). We infer that $E \rightarrow E - En^i \partial_i T$ so that

$$\delta f \longrightarrow \delta f + T\bar{f}' + En^i \partial_i T \frac{\partial \bar{f}}{\partial E}.$$

The unperturbed Boltzmann (6.110), i.e. $L[\bar{f}] = 0$, can be used to express $\bar{f}' = \mathcal{H}\bar{E}\partial\bar{f}/\partial E$ so that we recover the previous result (6.156).

We can therefore define a gauge invariant distribution function as

$$F \equiv \delta f - [C + n^i \partial_i (E' - B)] E \frac{\partial \bar{f}}{\partial E},$$

(6.157)

where we stress here that E refers to the metric perturbation in the term $(E' - B)$ and otherwise to the energy. The density contrast deduced from the expression (6.157) is

$$\int PE^3 dE d^2n = \delta\rho_\gamma + 4C = \rho_\gamma \delta_\gamma^P.$$

²Be careful here since T does not correspond to the temperature but to the $\nu = 0$ component of the vector field ξ^μ , as defined in Chapter 5.

Thus, F corresponds to the perturbation of the distribution function in the flat-slicing gauge. We could have defined another gauge invariant distribution function by

$$F = \mathcal{F} + \Psi E \frac{\partial \bar{f}}{\partial E} = \mathcal{F} + \bar{f}' \frac{\Psi}{\mathcal{H}}. \quad (6.158)$$

\mathcal{F} corresponds to the distribution function in the Newtonian gauge. After integrating over the energy, we obtain that $\delta I_F = \delta I_{\mathcal{F}} - 4\Psi\rho_{\gamma}$. The associated temperature contrasts, defined in (6.129), will be denoted by \mathcal{M} (flat-slicing gauge and obtained from F) and Θ (Newtonian gauge and obtained from \mathcal{F}). They are related by the relation

$$\mathcal{M} = \Theta - \Psi,$$

and satisfy the equations

$$\begin{aligned} \mathcal{M}' + n^k \partial_k \mathcal{M} - {}^{(3)}\Gamma_{ij}^k n^i n^j \frac{\partial \mathcal{M}}{\partial n^k} &+ \left[n^k \partial_k (\Phi + \Psi) + \left(\bar{E}'_{ij} + D_{(i} \bar{\Phi}_{j)} \right) n^i n^j \right], \\ &= C[\mathcal{M}] \end{aligned} \quad (6.159)$$

$$\begin{aligned} \Theta' + n^k \partial_k (\Theta + \Phi) - {}^{(3)}\Gamma_{ij}^k n^i n^j \frac{\partial \Theta}{\partial n^k} &+ \left[-\Psi' + \left(\bar{E}'_{ij} + D_{(i} \bar{\Phi}_{j)} \right) n^i n^j \right] \\ &= C[\Theta]. \end{aligned} \quad (6.160)$$

Since the metric terms do not depend on n , these two equations are identical, apart from the monopole and the dipole.

6.3.2.2 Boltzmann equation for the temperature

We restrict ourselves temporarily to a space-time with flat spatial sections and focus on the scalar modes. In Fourier space, the Boltzmann (6.160) for the temperature takes the form

$$\Theta' + ik\mu(\Theta + \Phi) = \Psi' + \tau' \left(\Theta_0 - \Theta + ik\mu V_b + \frac{1}{16} n_i n_j \Pi^{ij} \right), \quad (6.161)$$

with $\mu = \mathbf{k} \cdot \mathbf{n}/k$. We can decompose the radial part of $\Theta(k, \eta, \mu)$ into Legendre polynomials as

$$\Theta(k, \eta, \mu) = \sum_{\ell} (-i)^{\ell} \Theta_{\ell}(k, \eta) P_{\ell}(\mu) e^{+i\mathbf{k} \cdot \mathbf{x}}. \quad (6.162)$$

Using the property of the Legendre polynomials

$$(\ell + 1)P_{\ell+1}(\mu) = (2\ell + 1)\mu P_{\ell} - \ell P_{\ell-1},$$

the term $ik\mu\Theta$ is rewritten as

$$k \sum_{\ell} \left(\frac{\ell + 1}{2\ell + 3} \Theta_{\ell+1} - \frac{\ell}{2\ell - 1} \Theta_{\ell-1} \right) (-i)^{\ell} P_{\ell}(\mu).$$

Then noticing that $\Psi' + \tau' \Theta_0 = (\Psi' + \tau' \Theta_0) P_0$, $i\tau' k\mu V_b = -kV_b(-i)^1 P_1$ and, from (6.132), $\Pi^{ij} n_i n_j / 16 = (-i)^2 P_2 \Theta_2 / 10$, the Boltzmann equation can be decomposed into the hierarchy

$$\Theta'_0 = -\frac{k}{3}\Theta_1 + \Psi', \quad (6.163)$$

$$\Theta'_1 = k(\Theta_0 - \frac{2}{5}\Theta_2 + \Phi) - \tau'(kV_b + \Theta_1), \quad (6.164)$$

$$\Theta'_2 = k\left(\frac{2}{3}\Theta_1 - \frac{3}{7}\Theta_3\right) - \frac{9}{10}\tau'\Theta_2, \quad (6.165)$$

$$\Theta'_\ell = k\left(\frac{\ell}{2\ell-1}\Theta_{\ell-1} - \frac{\ell+1}{2\ell+3}\Theta_{\ell+1}\right) - \tau'\Theta_\ell. \quad (6.166)$$

As expected, the perturbations of the metric and of the baryon energy-momentum are only involved in the equations for the monopole and the dipole. Moreover, for $\ell > 2$, the collision term reduces to $-\tau'\Theta_\ell$, as expected.

With these conventions, the first multipoles of Θ are related to the fluid quantities by

$$4\Theta_0 = \delta_\gamma^N, \quad \Theta_1 = -kV_\gamma, \quad \Theta_2 = \frac{5}{12}k^2\pi_\gamma. \quad (6.167)$$

The factor $-k$ between Θ_1 and V_γ is a consequence of the relation (6.141). The first two equations of the hierarchy are therefore only the continuity and Euler equations for radiation, as always in kinetic theory.

Now, for the tensor modes, the Boltzmann equation reduces to

$$\Theta' + ik\mu\Theta = -\bar{E}'_{ij}n^i n^j - \tau'\left(\Theta - \frac{1}{16}n_i n_j \Pi^{ij}\right). \quad (6.168)$$

Decomposing \bar{E}_{ij} as $\bar{E}_{ij} = H^{(\pm 2)}Q_{ij}^{(\pm 2)}$ [see (6.141)], the hierarchy of the tensor part starts at $\ell = 2$, namely

$$\Theta'_2 = -k\frac{\sqrt{5}}{7}\Theta_2 - \frac{9}{10}\tau'\Theta_2 - \dot{H}, \quad (6.169)$$

$$\Theta'_\ell = k\left(\frac{\sqrt{\ell^2-4}}{2\ell-1}\Theta_{\ell-1} - \frac{\sqrt{(\ell+1)^2-4}}{2\ell+3}\Theta_{\ell+1}\right) - \tau'\Theta_\ell. \quad (6.170)$$

6.3.3 Thomson scattering and polarization

In the previous section, we have assumed that radiation was not polarized. However, Thomson scattering tends to polarize radiation in the direction perpendicular to the scattering plane, in particular if the radiation is anisotropic before its scattering. We should therefore take this polarization into account and determine its influence on the collision term. For complementary studies on polarization, we recommend Refs. [22–26]. The multipole approach is detailed in Refs. [3, 27].

6.3.3.1 Stokes parameters

The state of polarization of an electromagnetic wave is described in terms of the Stokes parameters [28]. The electric field of any monochromatic electromagnetic plane wave that propagates along the direction e_3 can be decomposed as

$$\mathbf{E}(x, t) = \mathcal{E} \exp[i(\omega t - \mathbf{k} \cdot \mathbf{x})], \quad \mathcal{E} \equiv \begin{pmatrix} a_1 \exp[i\theta_1] \\ a_2 \exp[i\theta_2] \\ 0 \end{pmatrix}. \quad (6.171)$$

If $\theta_1 = \theta_2$, E_1 and E_2 have the same phase and the wave is polarized linearly. Its polarization vector then forms an angle such that $\tan \alpha = E_2/E_1$ with e_1 . If E_1 and E_2 have different phases, the wave is polarized elliptically. Instead of the two amplitudes, a_1 and a_2 , and the two phases θ_1 and θ_2 , we can describe this wave with the four Stokes parameters,

$$I \equiv \langle a_1^2 \rangle + \langle a_2^2 \rangle = I_1 + I_2, \quad (6.172)$$

$$Q \equiv \langle a_1^2 \rangle - \langle a_2^2 \rangle = I_1 - I_2, \quad (6.173)$$

$$U \equiv 2\langle a_1 a_2 \cos(\theta_1 - \theta_2) \rangle, \quad (6.174)$$

$$V \equiv 2\langle a_1 a_2 \sin(\theta_1 - \theta_2) \rangle. \quad (6.175)$$

Here the averages are taken over several periods, assuming that the amplitudes and phases could vary in time, but slowly compared to the period $2\pi/\omega$ of the wave. Only three of these parameters are independent since they clearly satisfy the relation

$$I^2 = Q^2 + U^2 + V^2. \quad (6.176)$$

I is the total intensity of the wave and the parameter Q measures the prevalence of the linear polarization along e_1 compared to that along e_2 . U and V give information on the phases. Expanding the wave (6.171) in the basis $e^{(\pm)} = (e_1 \pm ie_2)/\sqrt{2}$, then $V = \langle a_+^2 \rangle - \langle a_-^2 \rangle$. Thus, V represents the difference between the positive and negative helicity intensities.

Under the action of a rotation of angle ψ of the plane (e_1, e_2) around the axis e_3 ,

$$e'_1 = \cos \psi e_1 + \sin \psi e_2, \quad e'_2 = -\sin \psi e_1 + \cos \psi e_2,$$

I and V remain invariant while Q and U transform as

$$\begin{pmatrix} Q' \\ U' \end{pmatrix} = \begin{pmatrix} \cos 2\psi & \sin 2\psi \\ -\sin 2\psi & \cos 2\psi \end{pmatrix} \begin{pmatrix} Q \\ U \end{pmatrix}. \quad (6.177)$$

In an equivalent way (I_1, I_2, U, V) transforms as

$$\begin{pmatrix} I'_1 \\ I'_2 \\ U' \\ V' \end{pmatrix} = \begin{pmatrix} \cos^2 \psi & \sin^2 \psi & (\sin 2\psi)/2 & 0 \\ \sin^2 \psi & \cos^2 \psi & -(\sin 2\psi)/2 & 0 \\ -\sin 2\psi & \sin 2\psi & \cos 2\psi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \\ U \\ V \end{pmatrix}. \quad (6.178)$$

From Q and U , we can thus construct two quantities that transform as a spin 2 object under the action of rotation

$$(Q \pm iU)'(e_3) = e^{\mp 2i\psi} (Q \pm iU)(e_3). \quad (6.179)$$

Since Q and U vanish in the unperturbed space-time, they are first-order (gauge invariant) quantities and they evolve without coupling to the metric perturbations (since such terms would be of order 2). Their Boltzmann equations then take the form

$$L \left[\begin{pmatrix} Q \\ U \end{pmatrix} \right] = \begin{pmatrix} Q \\ U \end{pmatrix}' + n^i \partial_i \begin{pmatrix} Q \\ U \end{pmatrix} - \mathcal{H}E \frac{\partial}{\partial E} \begin{pmatrix} Q \\ U \end{pmatrix} = C \left[\begin{pmatrix} Q \\ U \end{pmatrix} \right]. \quad (6.180)$$

Here, we have neglected the curvature term that could easily be added. By redefining the Stokes parameters as

$$\begin{pmatrix} Q \\ U \end{pmatrix} \rightarrow \frac{1}{4\bar{I}} \int 4\pi E^2 \begin{pmatrix} Q \\ U \end{pmatrix} dE, \quad (6.181)$$

and using the evolution (6.126) for the brightness \bar{I} , we obtain the Liouville equation for the polarization

$$L \left[\begin{pmatrix} Q \\ U \end{pmatrix} \right] = \begin{pmatrix} Q \\ U \end{pmatrix}' + n^i \partial_i \begin{pmatrix} Q \\ U \end{pmatrix} = C \left[\begin{pmatrix} Q \\ U \end{pmatrix} \right], \quad (6.182)$$

with

$$C \left[\begin{pmatrix} Q \\ U \end{pmatrix} \right] \rightarrow \frac{1}{4\bar{I}} \int 4\pi E^2 C \left[\begin{pmatrix} Q \\ U \end{pmatrix} \right] dE. \quad (6.183)$$

6.3.3.2 Spherical harmonics of spin s

For any direction e_3 defined by the two angles (θ, ϕ) , one can introduce a tangent basis on the sphere, (e_1, e_2) , defined up to a rotation. A function ${}_s f(\theta, \phi)$ defined on the sphere is then said to be of spin s if it transforms as ${}_s f'(\theta, \phi) = e^{-is\psi} {}_s f(\theta, \phi)$ under a rotation of angle ψ around the axis e_3 .

Any function of spin s on the sphere can be expanded in the basis of spherical harmonics of spin s [29], ${}_s Y_{\ell m}$, which satisfies the completeness and orthogonality relations

$$\int_0^{2\pi} d\phi \int_{-1}^1 d\cos\theta {}_s Y_{\ell'm'}^*(\theta, \phi) {}_s Y_{\ell m}(\theta, \phi) = \delta_{\ell\ell'} \delta_{mm'}, \quad (6.184)$$

$$\sum_{\ell m} {}_s Y_{\ell m}^*(\theta, \phi) {}_s Y_{\ell m}(\theta', \phi') = \delta(\phi - \phi') \delta(\cos\theta - \cos\theta'). \quad (6.185)$$

A central property in the construction of these spherical harmonics is the existence of operators that can be used to increase (∂^+) or decrease (∂^-) the spin of any function,

in the sense where $\partial^{\pm}{}_s f$ transforms as $(\partial^{\pm}{}_s f)' = e^{-i(s\pm 1)\psi} \partial^{\pm}{}_s f$ under a rotation of angle ψ . The explicit expression for these operators is

$$\partial^{\pm}{}_s f(\theta, \phi) = -\sin^{\pm s} \theta \left(\partial_\theta \pm \frac{i}{\sin \theta} \partial_\phi \right) \sin^{\mp s} \theta {}_s f(\theta, \phi). \quad (6.186)$$

In particular, for functions of spin 2 that satisfy $\partial_\phi [\pm_2 f(\mu, \phi)] = im \pm_2 f(\mu, \phi)$, these operators reduce to

$$(\partial^{\mp})^2 \pm_2 f(\mu, \phi) = \left(-\partial_\mu \pm \frac{m}{1 - \mu^2} \right)^2 [(1 - \mu^2) \pm_2 f(\mu, \phi)], \quad (6.187)$$

with $\mu = \cos \theta$.

We can then define the spherical harmonics of spin s from the action of these operators on the (usual) spherical harmonics

$${}_s Y_{\ell m} \equiv \sqrt{\frac{(\ell - s)!}{(\ell + s)!}} (\partial^+)^s Y_{\ell m}, \quad {}_s Y_{\ell m} \equiv \sqrt{\frac{(\ell + s)!}{(\ell - s)!}} (-1)^s (\partial^-)^{-s} Y_{\ell m}, \quad (6.188)$$

respectively, for $0 \leq s \leq \ell$ and $-\ell \leq s \leq 0$. These functions satisfy the following conditions

$${}_s Y_{\ell m}^* = (-1)^s {}_{-s} Y_{\ell -m}, \quad (6.189)$$

$$\partial^{\pm} {}_s Y_{\ell m} = \pm \sqrt{(\ell \mp s)(\ell + 1 \pm s)} {}_{s \pm 1} Y_{\ell m}, \quad (6.190)$$

$$\partial^- \partial^+ {}_s Y_{\ell m} = -(\ell - s)(\ell + 1 + s) {}_s Y_{\ell m}. \quad (6.191)$$

6.3.3.3 E and B modes

Just as the temperature is expanded as

$$\Theta(\mathbf{n}) = \sum_{\ell m} a_{\ell m} Y_{\ell m}(\mathbf{n}), \quad (6.192)$$

the polarization field can be expanded as

$$(Q \pm iU)(\mathbf{n}) = \sum_{\ell m} a_{\ell m}^{(\pm 2)} \pm_2 Y_{\ell m}(\mathbf{n}) \quad (6.193)$$

with $a_{\ell m}^{(-2)*} = a_{\ell -m}^{(+2)}$. The relation (6.189) can be used to express this last expansion as

$$(\partial^{\pm})^2 (Q \pm iU)(\mathbf{n}) = \sqrt{\frac{(\ell + 2)!}{(\ell - 2)!}} \sum_{\ell m} a_{\ell m}^{(\pm 2)} Y_{\ell m}(\mathbf{n}). \quad (6.194)$$

So we have been able to construct rotational invariant quantities from the Stokes parameters at the price of the action of a non-local operator. The coefficients of these expansions are given by

$$a_{\ell m} = \int d^2 n Y_{\ell m}^*(\mathbf{n}) \Theta(\mathbf{n}), \quad (6.195)$$

and

$$a_{\ell m}^{(\pm 2)} = \int d^2 n \pm_2 Y_{\ell m}^*(Q \pm iU) = \sqrt{\frac{(\ell+2)!}{(\ell-2)!}} \int d^2 n Y_{\ell m}^*(n) (\partial^\mp)^2 (Q \pm iU)(n). \quad (6.196)$$

One of the difficulties introduced by this expansion comes from the fact that the Stokes parameters are not invariant under rotation. So if the latter are easily computed for a given mode \mathbf{k} , they must be recomputed again in a fixed frame when integrating over \mathbf{k} . To avoid this problem, it is convenient to decompose the polarization as

$$(Q \pm iU)(\eta, \mathbf{x}, \mathbf{n}) = - \sum_{\ell m} (E_{\ell m} \pm iB_{\ell m}) \pm_2 Y_{\ell m}(\mathbf{n}), \quad (6.197)$$

so that

$$E_{\ell m} = -\frac{1}{2} [a_{\ell m}^{(+2)} + a_{\ell m}^{(-2)}], \quad B_{\ell m} = -\frac{1}{2i} [a_{\ell m}^{(+2)} - a_{\ell m}^{(-2)}]. \quad (6.198)$$

The choice of the designation E and B is related to the properties of parity of these quantities. Under a reflection with respect to the plane perpendicular to \mathbf{k} , Q remains unchanged and U transforms into $-U$. The spherical harmonics of spin s transform into $(-1)^\ell {}_s Y_{\ell m}$ under this operation, which implies that $E_{\ell m}$ transforms into $(-1)^\ell E_{\ell m}$, while $B_{\ell m}$ transforms into $(-1)^{\ell+1} B_{\ell m}$. Thus, E represents the scalar (or electric) part of the polarization and B the pseudo-scalar (or magnetic) part. Figure 6.6 illustrates configurations of fields with pure E and B modes.

6.3.3.4 Relation with the Stokes parameters

To better understand the relation between the E and B modes and the Stokes parameters, it can be useful to work in the flat-sky limit in which the coordinates (ϑ, φ) on the sphere are replaced by polar coordinates (r, φ) with $r = 2 \sin(\vartheta/2)$. This implies that we can replace the indices (ℓ, m) by a bidimensional vector $\ell = (\ell_x, \ell_y)$ with azimuthal angle φ_ℓ so that $\ell_x + i\ell_y = \ell \exp(i\varphi_\ell)$.

In this limit, $Y_{\ell m}(n) \rightarrow \exp(i\ell \cdot \mathbf{n})$ and $\pm_2 Y_{\ell m}(n) \rightarrow \exp(i\ell \cdot \mathbf{n} \pm 2i\phi_\ell)$ so that the temperature and polarization fields can be expanded in bidimensional Fourier modes as

$$\Theta(\mathbf{n}) = \int \frac{d^2 \ell}{2\pi} \Theta(\ell) e^{i\ell \cdot \mathbf{n}}, \quad (6.199)$$

$$(Q \pm iU)(\mathbf{n}) = - \int \frac{d^2 \ell}{2\pi} [E(\ell) \pm iB(\ell)] \frac{1}{\ell^2} (\partial^\pm)^2 e^{i\ell \cdot \mathbf{n}}. \quad (6.200)$$

In the small angle limit, (6.186) implies that

$$(\partial^\pm)^2 e^{i\ell \cdot \mathbf{n}} = -\frac{(\ell_x \pm \ell_y)^2}{\ell^2} e^{\mp 2i\phi_\ell} e^{i\ell \cdot \mathbf{n}} = -\ell^2 e^{\pm 2i(\phi_\ell - \phi)} e^{i\ell \cdot \mathbf{n}}.$$

The previous expressions have been derived in the natural spherical basis (e_r, e_θ, e_ϕ) but in the small angle limit, one can define a fixed basis in the plane perpendicular

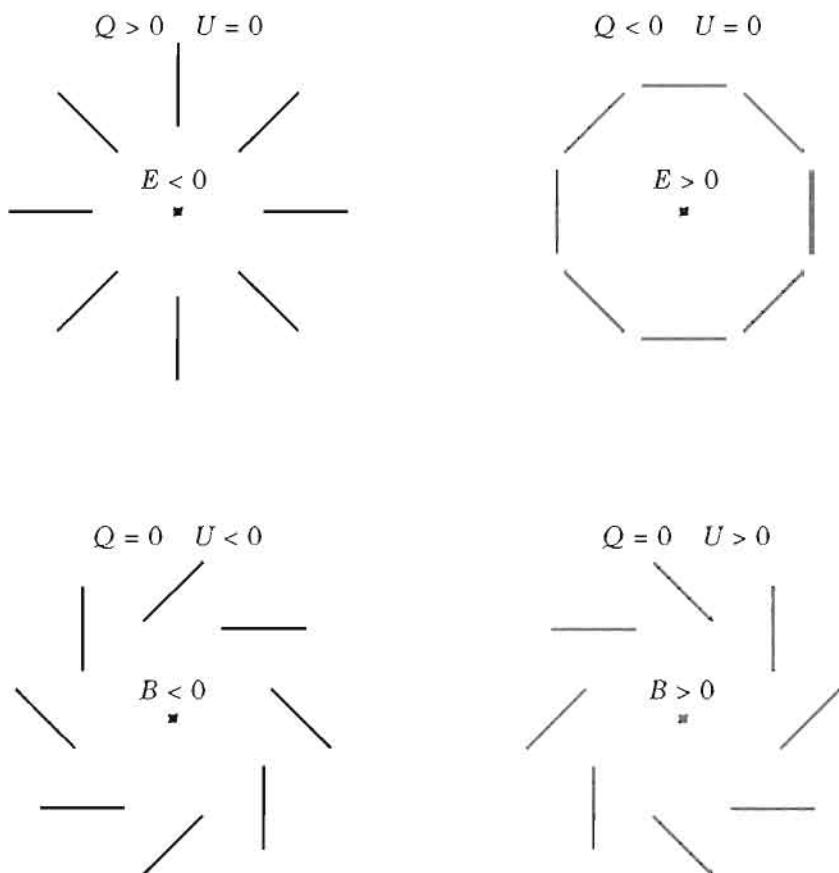


Fig. 6.6 The polarization figures correspond to pure E or B modes at a given point, chosen as the centre. The bars give the orientation of the polarization at the centre of the segment and thus give a simple representation of the Stokes parameters. These configurations repeat themselves identically to infinity. Since the components E and B are non-local, one should consider the configuration of the polarization field all around the point to determine these two components. We recover the properties of parity of both components, the top two configurations (E modes) remain unchanged under such a transformation, while the bottom ones (B modes) are exchanged.

to the local vertical, z . The Stokes parameters in the two systems of coordinates are related by $(Q \pm iU)' = e^{\mp 2i\phi}(Q \pm iU)$ so that, in this frame,

$$Q(n) = \int \frac{d^2\ell}{2\pi} [E(\ell) \cos 2\phi_\ell - B(\ell) \sin 2\phi_\ell] e^{i\ell \cdot n}, \quad (6.201)$$

$$U(n) = \int \frac{d^2\ell}{2\pi} [E(\ell) \sin 2\phi_\ell - B(\ell) \cos 2\phi_\ell] e^{i\ell \cdot n}. \quad (6.202)$$

Since

$$\cos 2\phi_\ell = (\ell_x^2 - \ell_y^2)/\ell^2, \quad \text{and} \quad \sin 2\phi_\ell = 2\ell_x \ell_y/\ell^2, \quad (6.203)$$

the previous expressions take the form

$$\Delta E = (\partial_x^2 - \partial_y^2)Q + 2\partial_x \partial_y U \quad \text{and} \quad \Delta B = (\partial_x^2 - \partial_y^2)U - 2\partial_x \partial_y Q, \quad (6.204)$$

in real space. The variables E and B are therefore non-local. This non-locality, inherited from (6.194), arises from the fact that the relation between the spherical harmonics of spin 0 and of spin 2 is also non-local.

6.3.3.5 Temperature and polarization expansion

The analysis of the previous sections shows that the radiation temperature $\Theta(\eta, x, n)$ at a point x propagating in the direction n can be expanded on the basis $G_{\ell m}(x, n)$ defined by

$$G_{\ell m}(x, n) = (-i)^\ell \sqrt{\frac{4\pi}{2\ell+1}} e^{ik \cdot x} Y_{\ell m}(n), \quad (6.205)$$

and that the polarization field can be expanded on the basis

$$\pm_2 G_{\ell m}(x, n) = (-i)^\ell \sqrt{\frac{4\pi}{2\ell+1}} e^{ik \cdot x} \pm_2 Y_{\ell m}(n). \quad (6.206)$$

We therefore expand the spatial part of Θ and $(Q \pm iU)$ on a basis of complex exponentials and their angular part on a basis of spherical harmonics of appropriate spin

$$\Theta(\eta, x, n) = \int \frac{d^3 k}{(2\pi)^{3/2}} \sum_{\ell m} \Theta_\ell^{(m)}(\eta, k) G_{\ell m}(x, n), \quad (6.207)$$

$$(Q \pm iU)(\eta, x, n) = \int \frac{d^3 k}{(2\pi)^{3/2}} \sum_{\ell m} [E_\ell^{(m)} \pm iB_\ell^{(m)}](\eta, k) \pm_2 G_{\ell m}(x, n). \quad (6.208)$$

The Stewart–Walker lemma implies that the variables $E_\ell^{(m)}$ and $B_\ell^{(n)}$ are gauge invariant.

6.3.3.6 Scattering cross-section

The Thomson scattering cross-section for an incident wave with linear polarization \mathcal{E}^{in} scattered to an outgoing wave with linear polarization \mathcal{E}^{out} is

$$\frac{d\sigma}{d\Omega} = \frac{3}{8\pi} \sigma_T |\mathcal{E}^{in*} \cdot \mathcal{E}^{out}|^2, \quad (6.209)$$

(see Ref. [30]). It is convenient to introduce the partial intensities

$$I_1 \equiv a_1^2 = \frac{I+Q}{2}, \quad I_2 \equiv a_2^2 = \frac{I-Q}{2}. \quad (6.210)$$

An incoming wave scattered at an angle θ in the plane (n, \mathbf{E}_2) (Fig. 6.7) has outgoing intensities

$$I_1^{\text{out}} = \frac{3\sigma_T}{8\pi} I_1^{\text{in}}, \quad I_2^{\text{out}} = \frac{3\sigma_T}{8\pi} \cos^2 \theta I_2^{\text{in}}, \quad (6.211)$$

which in terms of the Stokes parameters translates into

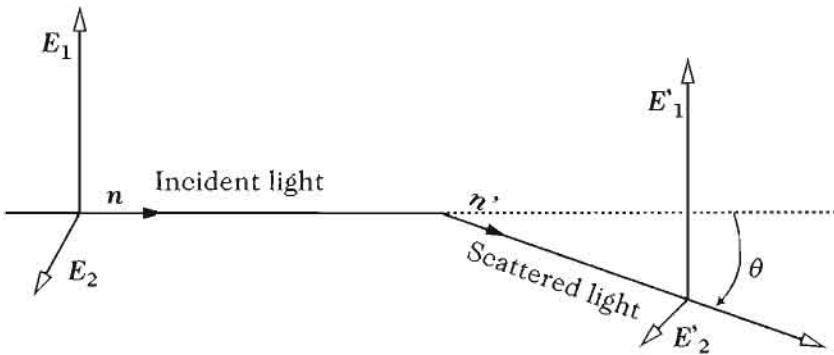


Fig. 6.7 Geometry of the Thomson scattering in the electron proper frame. An incident photon in the direction \mathbf{n} is deflected by an angle θ and comes out in the direction \mathbf{n}' .

$$\begin{pmatrix} I_1^{\text{out}} \\ Q^{\text{out}} \end{pmatrix} = \frac{3\sigma_T}{16\pi} \begin{pmatrix} 1 + \cos^2 \theta & \sin^2 \theta \\ \sin^2 \theta & 1 + \cos^2 \theta \end{pmatrix} \begin{pmatrix} I_1^{\text{in}} \\ Q^{\text{in}} \end{pmatrix}. \quad (6.212)$$

Under a rotation, (Q, U) transforms as (6.177). To determine the cross-section of an initial state $(I^{\text{in}}, Q^{\text{in}}, U^{\text{in}})$ propagating in the direction \mathbf{n} scattered into a final state $(I^{\text{out}}, Q^{\text{out}}, U^{\text{out}})$ that propagates in the direction \mathbf{n}' , we perform the following operations, using the plane (y, z) as reference plane, (see Fig. 6.8 for a definition of the angles and the vectors and Ref. [30] for details).

1. We perform a rotation around \mathbf{n} to bring the plane $(\mathbf{n}, \mathbf{n}')$ into the plane (\mathbf{n}, \mathbf{z}) , which translates into a rotation (6.177) with $\phi = \alpha$ for the Stokes parameters.
2. We perform a rotation around the axis \mathbf{z} to bring the new plane $(\mathbf{n}, \mathbf{n}')$ into the reference plane (y, z) . This operation does not influence the Stokes parameters.
3. We can then use the relations (6.212).
4. We perform a rotation around the axis \mathbf{z} to bring the scattering plane back to the old plane (z, n') . Again, this operation does not affect the Stokes parameters.
5. Finally, we perform a rotation around \mathbf{n}' to recover the initial position, which translates into a rotation (6.177) with $\phi = \alpha'$ for the Stokes parameters.

Omitting the parameter V that decouples from the other polarizations and remains zero if it was initially, the previous procedure allows us to relate the outgoing Stokes parameters to their incoming values

$$\begin{pmatrix} I_1^{\text{out}} \\ I_2^{\text{out}} \\ U^{\text{out}} \end{pmatrix} = \frac{3\sigma_T}{4\pi} \mathcal{P} \begin{pmatrix} I_1^{\text{in}} \\ I_2^{\text{in}} \\ U^{\text{in}} \end{pmatrix}. \quad (6.213)$$

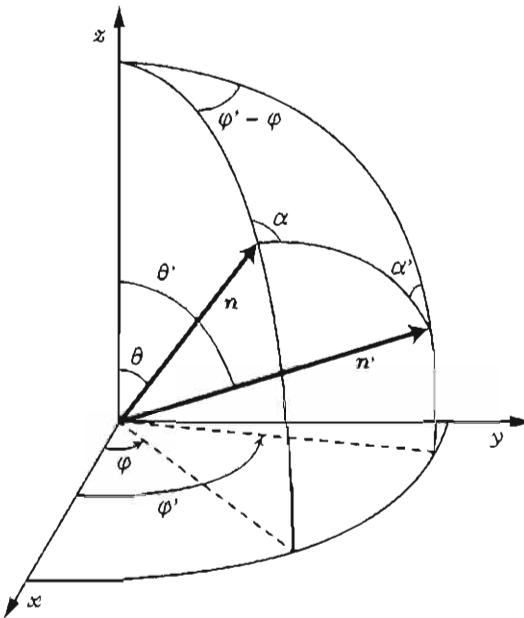


Fig. 6.8 Definition of the angles and vectors involved in Thomson scattering.

The matrix \mathcal{P} can be decomposed as

$$\mathcal{P} = \mathcal{P}^{(0)} + \sqrt{(1-\mu^2)(1-\mu'^2)}\mathcal{P}^{(1)} + \mathcal{P}^{(2)}, \quad (6.214)$$

with

$$\mathcal{P}^{(0)} = \frac{3}{4} \begin{pmatrix} 2(1-\mu^2)(1-\mu'^2) + \mu^2\mu'^2 & \mu^2 & 0 \\ \mu'^2 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad (6.215)$$

$$\mathcal{P}^{(1)} = \frac{3}{4} \begin{pmatrix} 4\mu\mu' \cos\delta & 0 & 2\mu\sin\delta \\ 0 & 0 & 0 \\ -4\mu'\sin\delta & 0 & 2\cos\delta \end{pmatrix}, \quad (6.216)$$

$$\mathcal{P}^{(2)} = \frac{3}{4} \begin{pmatrix} \mu^2\mu'^2 \cos 2\delta & -\mu^2 \cos 2\delta & \mu^2\mu' \sin 2\delta \\ -\mu'^2 \cos 2\delta & \cos 2\delta & -\mu' \sin 2\delta \\ -2\mu\mu'^2 \sin 2\delta & 2\mu\sin 2\delta & 2\mu\mu' \cos 2\delta \end{pmatrix}, \quad (6.217)$$

with $\delta = \varphi' - \varphi$. We then have access to the Stokes parameters by performing the rotation

$$\begin{pmatrix} I_1 \\ I_2 \\ U \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} I \\ Q \\ U \end{pmatrix}. \quad (6.218)$$

For each polarization, the collision term is given, in the rest-frame of the baryons, by

$$C^{(\lambda)} = \sigma_T n_e \left[\int f^{(\lambda')}(E', n') \mathcal{P}_\lambda^\lambda(n, n') \frac{d^2 n'}{4\pi} - f^{(\lambda)}(E, n) \right], \quad (6.219)$$

where $\mathcal{P}_{\lambda'}^{\lambda}(\mathbf{n}, \mathbf{n}')$ corresponds to the 2×2 block of \mathcal{P} corresponding to (I_1, I_2) . This expression generalizes (6.144).

This collision term can be expanded on the harmonics basis (6.205) and (6.206). For that, it is convenient to introduce the vector $\mathbf{T} \equiv (\Theta, Q + iU, Q - iU)$. It can then be shown (which we do not demonstrate here, see Ref. [31]) that the collision term takes the form

$$C[\mathbf{T}] = \tau' \left[-\mathbf{T} + (\Theta_0 + n_i V_b^i) + \frac{1}{10} \sum_{m=-2}^2 \int \mathbf{P}^{(m)}(\mathbf{n}, \mathbf{n}') \mathbf{T}(\mathbf{n}') d^2 n' \right]. \quad (6.220)$$

The term $\mathbf{P}^{(m)}$ is deduced from (6.214)–(6.218) and takes the form

$$\mathbf{P}^{(m)} = \begin{pmatrix} Y_{2m}(\mathbf{n}') Y_{2m}(\mathbf{n}) & -\sqrt{\frac{3}{2}} {}_2Y_{2m}(\mathbf{n}') Y_{2m}(\mathbf{n}) & -\sqrt{\frac{3}{2}} {}_{-2}Y_{2m}(\mathbf{n}') Y_{2m}(\mathbf{n}) \\ -\sqrt{6} Y_{2m}(\mathbf{n}') {}_2Y_{2m}(\mathbf{n}) & {}_2Y_{2m}(\mathbf{n}') {}_2Y_{2m}(\mathbf{n}) & {}_{-2}Y_{2m}(\mathbf{n}') {}_2Y_{2m}(\mathbf{n}) \\ -\sqrt{6} Y_{2m}(\mathbf{n}') {}_2Y_{2m}(\mathbf{n}) & {}_2Y_{2m}(\mathbf{n}') {}_{-2}Y_{2m}(\mathbf{n}) & {}_{-2}Y_{2m}(\mathbf{n}') {}_{-2}Y_{2m}(\mathbf{n}) \end{pmatrix}. \quad (6.221)$$

Note that if one restricts attention to the temperature, this expression reduces to (6.151).

6.3.3.7 Origin of the cosmic microwave background polarization

The properties of the Thomson scattering that we have just described allow us to understand from a heuristic point of view the origin of the cosmic microwave background polarization (Fig. 6.9).

A mode propagating along the direction x has a polarization in the plane (y, z) . If an initially unpolarized wave is scattered in the direction z , the outgoing wave ends up being linearly polarized in the direction y . However, if we consider incident waves coming from every direction, then statistical isotropy implies that the outgoing wave is not polarized (on average), see Fig. 6.9a.

If initially radiation has a dipolar anisotropy, the polarization along y of the waves propagating in the directions x and $-x$ has on average the same amplitude as the polarization along x of the waves propagating in the directions y so that the outgoing wave along z is, again, not polarized (on average), see Fig. 6.9b.

To produce a polarized outgoing wave, the radiation must have a non-vanishing quadrupole (see Fig. 6.9c). The hot and cold zones are then in orthogonal directions compared to the scattering point.

This implies that the polarization of the cosmic microwave background will be weaker than the temperature anisotropies. Indeed, the Thomson scattering, which sources this polarization, is only efficient when the Universe is ionized. At this epoch, photons and baryons are strongly coupled so that the quadrupole of the radiation is very weak. For scalar modes, the polarization amplitude on super-Hubble scales (i.e. for multipoles lower than the position of the first Doppler peak) must therefore be strongly damped. The polarization field will in addition exhibit the same acoustic peaks as those

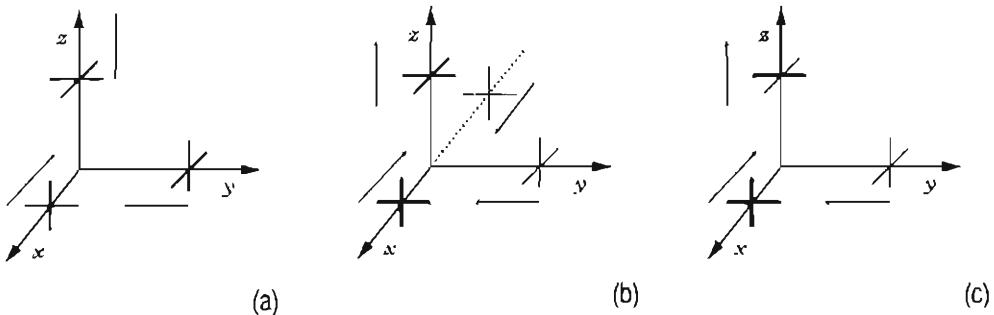


Fig. 6.9 (a) Isotropy implies that the Thomson scattering does not generate any polarization. The polarization amplitude of the waves propagating in the directions x and y are equal. The polarization of the outgoing wave carries away the y -component of the incoming wave that propagates along x and the x -component of the incoming wave propagating along y , which have the same amplitude. (b) Case of dipolar anisotropy: the polarization of the outgoing wave is the sum of the y -components of the incoming waves propagating along x (coming from a hot zone) and $-x$ (coming from a cold zone) that, on average, has the same amplitude as the x -components of the incoming wave propagating along y . It is thus also unpolarized. (c) Case of a quadrupole anisotropy: the outgoing wave has a polarization along y , inherited from the incoming wave along x that comes from a hot region, larger than that along x , inherited from the incoming wave along y that comes from a cold region.

for the temperature spectrum but, due to the quadrupole origin, the position of the peaks will be out-of-phase. On the other hand, the primordial gravitational waves that intrinsically have a quadrupolar anisotropy, even on super-Hubble scales, can generate polarization fields, even at large angular scales.

6.3.3.8 Temperature and polarization hierarchy

Choosing the direction e_3 along the direction k allows us to express the gradient term as

$$n^i \partial_i \rightarrow i n^i k_i = i \sqrt{\frac{4\pi}{3}} k Y_{10}(n).$$

This term will multiply the angular dependence of the temperature and the polarization. This introduces terms of the form $Y_{10}Y_{\ell m}$ and $Y_{10\pm 2}Y_{\ell m}$ that are expressed as

$$\begin{aligned} \sqrt{\frac{4\pi}{3}} Y_{10} {}_s Y_{\ell m} &= \frac{{}_s \kappa_{\ell}^m}{\sqrt{(2\ell+1)(2\ell-1)}} {}_s Y_{\ell-1 m} - \frac{ms}{\ell(\ell+1)} {}_s Y_{\ell m} \\ &+ \frac{{}_s \kappa_{\ell+1}^m}{\sqrt{(2\ell+1)(2\ell+3)}} {}_s Y_{\ell+1 m}, \end{aligned} \quad (6.222)$$

with

$${}_s \kappa_{\ell m}^{\ell} \equiv \sqrt{\frac{(\ell^2 - m^2)(\ell^2 - s^2)}{\ell^2}}. \quad (6.223)$$

The explicit form of the Boltzmann equation follows directly from this decomposition to give

$$\Theta_\ell^{(m)'} = k \left[\frac{0\kappa_\ell^m}{(2\ell-1)} \Theta_{\ell-1}^{(n)} - \frac{0\kappa_{\ell+1}^m}{(2\ell+3)} \Theta_{\ell+1}^{(m)} \right] - \tau' \Theta_\ell^{(m)} + S_\ell^{(n)}, \quad (6.224)$$

$$\begin{aligned} E_\ell^{(m)'} = k & \left[\frac{2\kappa_\ell^m}{(2\ell-1)} E_{\ell-1}^{(m)} - \frac{2m}{\ell(\ell+1)} B_\ell^{(m)} - \frac{2\kappa_{\ell+1}^m}{(2\ell+3)} E_{\ell+1}^{(m)} \right] \\ & - \tau' \left[E_\ell^{(m)} + \sqrt{6} P^{(m)} \delta_{\ell,2} \right], \end{aligned} \quad (6.225)$$

$$\begin{aligned} B_\ell^{(m)'} = k & \left[\frac{2\kappa_\ell^m}{(2\ell-1)} B_{\ell-1}^{(m)} + \frac{2m}{\ell(\ell+1)} E_\ell^{(m)} - \frac{2\kappa_{\ell+1}^m}{(2\ell+3)} B_{\ell+1}^{(m)} \right] \\ & - \tau' B_\ell^{(m)}. \end{aligned} \quad (6.226)$$

The source terms are given by

$$\begin{aligned} S_0^{(0)} &= \tau' \Theta_0^{(0)} + \Psi, & S_1^{(0)} &= -k\tau' V_b + k\Phi, & S_2^{(0)} &= \tau' P^{(0)}, \\ S_1^{(1)} &= -\tau' V_b^{(1)}, & S_2^{(1)} &= \tau' P^{(1)} + \frac{k}{\sqrt{3}} \Phi^{(1)}, & & \\ S_2^{(2)} &= \tau' P^{(2)} - H^{(2)'}, & & & & \end{aligned} \quad (6.227)$$

if the Fourier coefficients of the vector and tensor perturbations are expanded as

$$\bar{\Phi}_i = -\frac{i}{\sqrt{2}} \sum_{m=\pm 1} e_i^\pm \Phi^{(m)}, \quad \bar{E}_{ij} = -\sqrt{\frac{3}{8}} \sum_{m=\pm 2} e_i^\pm \otimes e_j^\pm H^{(m)}. \quad (6.228)$$

The term $P^{(m)}$ that reflects the anisotropic nature of the Compton scattering is given by

$$P^{(m)} = \frac{1}{10} \left[\Theta_2^{(m)} - \sqrt{6} E_2^{(m)} \right], \quad (6.229)$$

and only includes the quadrupole of the temperature and of the polarization E .

Notice that the polarization source, $P^{(m)}$, only enters into the equation of evolution for the quadrupole of E . The equations for the temperature are to be compared with those [(6.163) and (6.166)] obtained by neglecting the polarization. It is not necessary to solve these equations for $|m|$ and $-|m|$. By symmetry, $B_\ell^{(0)} = 0$, for any source. This set of equations replaces the fluid equations of evolution for the radiation. The other equations remain unchanged. For neutrinos, we can use similar equations with $\tau' = 0$.

Notice that in the short-wavelength limit ($k \gg \tau'$), the evolution equations for a scalar mode imply that

$$E_2^{(0)} = -\sqrt{6} P^{(0)}, \quad P^{(0)} = \frac{1}{4} \Theta_1^{(0)}, \quad \sqrt{6} \Theta_2^{(0)} = -4 E_2^{(0)},$$

and

$$\Theta_2^{(0)} = \frac{8}{9} \frac{k}{\tau'} \Theta_1^{(0)}. \quad (6.230)$$

Using these relations (6.167), we obtain

$$\frac{k^2}{12}\pi_\gamma = \frac{1}{5}\Theta_2^{(0)} = -\frac{8}{45}\frac{k^2}{\tau'}V_\gamma, \quad (6.231)$$

which is the relation we have used in the study of the Silk damping [see Section 6.2.3].

6.3.3.9 Power spectra

With the previous expansions, the power spectrum is given by

$$(2\ell+1)^2 C_\ell^{XZ} = \frac{2}{\pi} \int k^2 dk \sum_{m=-2}^{m=2} X_\ell^{(m)*}(\eta_0, k) Z_\ell^{(m)}(\eta_0, k), \quad (6.232)$$

where X and Z take the values Θ , E or B . Figure 6.10 illustrates the four non-vanishing spectra that can be computed.

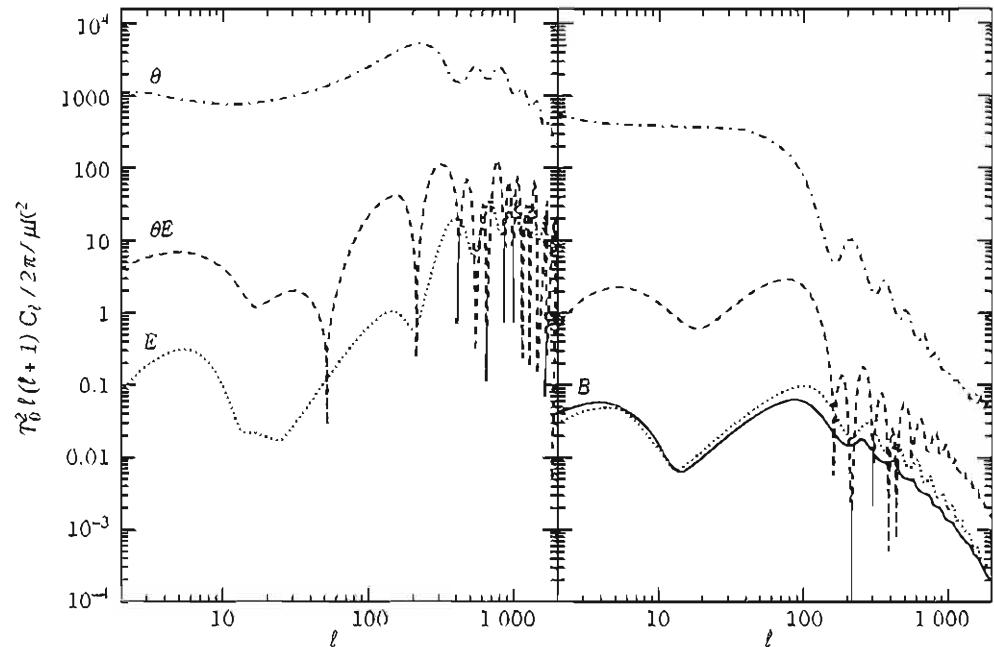


Fig. 6.10 The four spectra $\Theta - \Theta$, $\Theta - E$, $E - E$ and $B - B$ induced by scalar (left) and tensor (right) modes for an inflationary model. As stressed earlier, the spectrum $B - B$ is only generated by tensor modes and the $E - E$ spectrum vanishes by symmetry.

The numerical integration of the Boltzmann hierarchy can in principle be used to determine the power spectra. Historically, this was the first method used to solve this problem and it amounts to evolving the photon distribution function during the entire history of the Universe. Technically, this approach is not very good as one has to solve

a large number of coupled differential equations, typically of the order of $5\ell_{\max}$ for a large number of values k (typically of the order of a thousand in an interval $k\eta_0 \in [\ell_{\min}/10, 2\ell_{\max}]$). One should also carefully control the way the hierarchy is truncated to avoid any (unphysical) reflection of the power towards the lower multipoles.

6.3.4 Numerical integration

6.3.4.1 Radial modes

It is useful to rewrite the expression for the functions ${}_sG_{\ell m}$ by expanding the complex exponentials into spherical harmonics thanks to the relation (B.20)

$$\exp(ik \cdot x) = \sum_{\ell} (-i)^{\ell} \sqrt{4\pi(2\ell+1)} j_{\ell}(kr) Y_{\ell 0}(\mathbf{n}),$$

with $e_3 = k/k$ and $x = -rn$. The sign convention is such that \mathbf{n} corresponds to the direction of propagation of the radiation toward the observer, i.e. opposite to the direction of observation. Using the recursion relations (B.46) between the spherical Bessel functions, we can rewrite the functions ${}_sG_{\ell m}$ as

$$G_{\ell' m} = \sum_{\ell} (-i)^{\ell} \sqrt{4\pi(2\ell+1)} j_{\ell}^{(\ell' m)}(kr) Y_{\ell m}(\mathbf{n}), \quad (6.233)$$

and

$$\pm_2 G_{2m} = \sum_{\ell} (-i)^{\ell} \sqrt{4\pi(2\ell+1)} \left[\epsilon_{\ell}^{(m)}(kr) \pm i\beta_{\ell}^{(m)}(kr) \right] \pm_2 Y_{\ell m}(\mathbf{n}). \quad (6.234)$$

The functions $j_{\ell}^{(\ell' m)}$, $\epsilon_{\ell}^{(m)}$ and $\beta_{\ell}^{(m)}$ are then explicitly given by

$$\begin{aligned} j_{\ell}^{(00)} &= j_{\ell}, & j_{\ell}^{(10)} &= j_{\ell}', & j_{\ell}^{(20)} &= (3j_{\ell}'' + j_{\ell})/2, \\ j_{\ell}^{(11)} &= \sqrt{\frac{\ell(\ell+1)}{2}} \frac{j_{\ell}}{x}, & j_{\ell}^{(21)} &= \sqrt{\frac{3\ell(\ell+1)}{2}} \left(\frac{j_{\ell}}{x} \right)', \\ j_{\ell}^{(22)} &= \sqrt{\frac{3}{8}} \frac{(\ell+2)!}{(\ell-2)!} \frac{j_{\ell}}{x^2} \end{aligned} \quad (6.235)$$

and

$$\epsilon_{\ell}^{(0)} = j_{\ell}^{(22)}, \quad (6.236)$$

$$\epsilon_{\ell}^{(1)} = \frac{\sqrt{(\ell-1)(\ell+2)}}{2} \left(\frac{j_{\ell}}{x^2} + \frac{j_{\ell}'}{x} \right), \quad (6.237)$$

$$\epsilon_{\ell}^{(2)} = \frac{1}{4} \left(-j_{\ell} + j_{\ell}'' + 2\frac{j_{\ell}}{x^2} + 4\frac{j_{\ell}'}{x} \right), \quad (6.238)$$

$$\beta_{\ell}^{(0)} = 0, \quad (6.239)$$

$$\beta_{\ell}^{(1)} = \frac{\sqrt{(\ell-1)(\ell+2)}}{2} \frac{j_{\ell}}{x}, \quad (6.240)$$

$$\beta_{\ell}^{(2)} = \frac{1}{2} \left(j_{\ell}' + 2\frac{j_{\ell}}{x} \right). \quad (6.241)$$

Moreover, these functions satisfy $\epsilon_{\ell}^{(-m)} = \epsilon_{\ell}^{(m)}$ and $\beta_{\ell}^{(-m)} = -\beta_{\ell}^{(m)}$.

6.3.4.2 Integral solution

The Boltzmann equations (6.224)–(6.226) have a formal solution in an integral form. Using the expansion (6.233) and (6.234) for the functions $G_{\ell m}$ into radial and angular parts, we can show that the solution of the temperature evolution (6.224) is given by

$$\Theta_\ell^{(m)}(\eta_0, k) = (2\ell + 1) \int_0^{\eta_0} d\eta e^{-\tau} \sum_{\ell'} S_{\ell'}^{(m)}(\eta) j_\ell^{(\ell' m)}[k(\eta_0 - \eta)]. \quad (6.242)$$

The solution for the polarization takes the similar form

$$\left[E_\ell^{(m)} \pm i B_\ell^{(m)} \right](\eta_0, k) = -(2\ell + 1) \sqrt{6} \int_0^{\eta_0} d\eta \tau' e^{-\tau} P^{(m)}(\eta) \left[\epsilon_\ell^{(m)} \pm i \beta_\ell^{(m)} \right] [k(\eta_0 - \eta)]. \quad (6.243)$$

This clever method [32,33] is actually a generalization of the Sachs–Wolfe formula since it is equivalent to looking at the temperature of the emitting zones and integrating along the line of sight. Its huge advantage lies in the fact that it is sufficient to compute precisely only the moments of order 2 of the perturbed distribution function by solving the Boltzmann equation since the higher order moments do not enter into the source terms $S_\ell^{(m)}$. In practice, one can simply truncate the Boltzmann hierarchy at $\ell = 2$ even though it is preferable to go up to $\ell \sim 10$. The other multipoles are deduced from the integral expressions. This step is quick since the Bessel functions can be pre-computed.

The form (6.242) is similar to the expressions we had obtained in Section 6.1.2, which did not contain the radiation anisotropic stress since Thomson scattering was not included at this stage. In particular, we have seen how the functions $j_\ell^{(00)}$ and $j_\ell^{(10)}$ for the scalar modes and $j_\ell^{(11)}$ for the vector modes arose naturally when the appropriate basis was used for the expansion. We notice a difference of normalization between $j_\ell^{(22)}$ and $\tilde{j}_\ell^{(22)}$ which that due to the difference between the basis used for the expansion of the tensor part [see (6.143)].

6.3.4.3 Freely available codes

Several codes integrating the Boltzmann hierarchy for the temperature and the polarization numerically are freely available: CMBFAST [34], CAMB [35], CMBEASY [36] and COSMICS [37]. These codes use the advantages from the integration technique along the line of sight. They either use a synchronous gauge or the Newtonian gauge and are written in FORTRAN or C++. The freely available versions integrate the perturbation equations in the simplest case of a Λ CDM model; all extensions (dark energy, ...) can be included at will to adapt to one's favourite model.

6.4 Anisotropies of the cosmic microwave background

6.4.0.4 Observations of the cosmic microwave background anisotropies

We observe the cosmic microwave background on a sphere around us. A map of the cosmic microwave background can thus always be expanded into spherical harmonics to extract the coefficients $a_{\ell m}^{\text{obs}}$,

$$\Theta_{\text{obs}}(e) = \sum_{\ell m} a_{\ell m}^{\text{obs}} Y_{\ell m}(e). \quad (6.244)$$

The isotropy of the Universe implies that the angular two-point correlation function of the observed temperature fluctuations,

$$C_{\text{obs}}(\vartheta) \equiv \langle \Theta_{\text{obs}}(e) \Theta_{\text{obs}}(e') \rangle_{e \cdot e' = \cos \vartheta}, \quad (6.245)$$

can be obtained as an average on the sky. This function only depends on $\cos \vartheta$ and therefore can be expanded in Legendre polynomials as

$$C_{\text{obs}}(\vartheta) = \frac{1}{4\pi} \sum_{\ell} C_{\ell}^{\text{obs}} P_{\ell}(\cos \vartheta), \quad (6.246)$$

so that

$$C_{\ell}^{\text{obs}} = \frac{1}{2\ell + 1} \sum_m |a_{\ell m}^{\text{obs}}|^2. \quad (6.247)$$

The theoretical predictions can only be statistical. Therefore, the aim cannot be to compare the observed cosmic microwave background map to a given simulated map but to compare the statistical properties of the two temperature distributions (observed and theoretically predicted in a given model). The comparison between theory and observations therefore introduces the spatial analogy of an ergodic hypothesis since C_{ℓ}^{obs} is obtained from the sky average of a unique realization (our observable Universe), whereas C_{ℓ} is obtained from an ensemble average on some stochastic initial conditions in the frame of a model. This leads to the notion of cosmic variance.

6.4.0.5 Estimator of C_{ℓ} and cosmic variance

From a theoretical point of view, we have access to the angular power spectrum of the temperature field that is a homogeneous and isotropic random field. The coefficients $a_{\ell m}$ are also independent stochastic fields of vanishing mean value. This implies that

$$\langle a_{\ell m} \rangle = 0, \quad \langle a_{\ell m} a_{\ell' m'}^* \rangle = \delta_{\ell \ell'} \delta_{m m'} C_{\ell}.$$

The angular power spectrum C_{ℓ} is not directly observable but an estimator of C_{ℓ} can be obtained by summing over m as

$$\widehat{C}_{\ell} = \frac{1}{2\ell + 1} \sum_m a_{\ell m} a_{\ell m}^* = \frac{C_{\ell}}{2\ell + 1} V_{\ell}, \quad (6.248)$$

with the variable V_{ℓ} defined as

$$V_{\ell} \equiv \sum_m \frac{a_{\ell m} a_{\ell m}^*}{C_{\ell}}.$$

C_{ℓ}^{obs} , obtained by summing the $a_{\ell m}^{\text{obs}}$ over m , as in (6.248), thus represents a measure of the estimator \widehat{C}_{ℓ} .

In the case of inflationary models, the stochastic fields $a_{\ell m}$ have almost-Gaussian statistics, meaning that their one-point probability distribution function is

$$P(a_{\ell m}) = \frac{1}{\sqrt{2\pi}C_\ell} \exp\left(-\frac{a_{\ell m}^2}{2C_\ell}\right). \quad (6.249)$$

V_ℓ being the sum of the square of $2\ell + 1$ independent random Gaussian variables with the same variance, its probability distribution function follows a χ^2 distribution with $2\ell + 1$ degrees of freedom, i.e.

$$P_{\chi_i^2}(V_\ell) = \frac{V_\ell^{\ell/2}}{2^{\ell/2}\Gamma(\ell/2)} \exp\left(-\frac{V_\ell}{2}\right). \quad (6.250)$$

We infer the probability distribution function of \hat{C}_ℓ ,

$$P(\hat{C}_\ell) = \frac{\ell}{C_\ell} P_{\chi_i^2}\left(\frac{\ell\hat{C}_\ell}{C_\ell}\right). \quad (6.251)$$

When $\ell \rightarrow \infty$, the central limit theorem ensures that this distribution becomes Gaussian and that

$$\lim_{\ell \rightarrow \infty} \hat{C}_\ell = C_\ell. \quad (6.252)$$

The estimator is then said to be *coherent*. Since $\langle \hat{C}_\ell \rangle = C_\ell$, from (6.22), this estimator is not biased. Its variance can also be computed and we obtain

$$\langle \hat{C}_\ell^2 \rangle - \langle \hat{C}_\ell \rangle^2 = \frac{2}{2\ell + 1} C_\ell^2.$$

\hat{C}_ℓ is the best estimator of C_ℓ that we can construct since its variance is the smallest one can hope to obtain by estimating C_ℓ . The efficiency of the estimator is limited by the fact that, for a given multipole, there only exists $2\ell + 1$ independent modes and

$$\frac{\langle \hat{C}_\ell^2 \rangle - \langle \hat{C}_\ell \rangle^2}{C_\ell^2} = \frac{2}{2\ell + 1}. \quad (6.253)$$

This cosmic variance is an unavoidable statistical limitation and is mainly important at small multipoles.

To conclude, no statistical properties of the $a_{\ell m}$ are directly accessible from observations. To compare predictions to observations, one must construct estimators by summing over the $a_{\ell m}$. In the case where the primordial perturbations do not follow Gaussian statistics, the construction of a good estimator (unbiased and optimal) is much more difficult. In particular, the two-point correlation function does not completely characterize the distribution.

6.4.0.6 Status of the observations

The measurement of the cosmic microwave background anisotropies has undergone a revolution over the past decades and we refer to Refs. [38–40] for a detailed presentation of these observations.

The first revolution has been performed by the DMR experiment on board of the COBE satellite (COsmic Background Explorer) that has measured for the first time the temperature anisotropies of the cosmic microwave background up to a multipole of the order of $\ell \sim 20$, corresponding to an angular scale $\vartheta \gtrsim 7$ deg. As illustrated in Fig. 6.11, the state-of-the-art has quickly evolved over the past ten years. In 1997, the observations of COBE at small ℓ and the results of various ground-based experiments were not able to discriminate between two test models. From 1998, the results of the balloon experiments BOOMERanG and MAXIMA have made it possible to detect the first and then the second peak, which was then confirmed by the Archeops balloon. Since 2003, the observations of the WMAP satellite give a good measure of the temperature angular power spectrum up to $\ell \sim 1000$. Figure 6.11 summarizes the complementarity of these experiments.

The cosmic microwave background observation experiences can be classified into three categories.

1. Ground-based experiments measure the small-scale temperature anisotropies, mainly by interferometry (DASI [41], CAT [42], VLA [43], CBI for instance). These observations must be performed at altitude or at very dry sites. Their main problem lies in the atmospherical fluctuations.
2. Balloon experiments are carried onboard balloons flying at an altitude of around 40 km, which reduces the problems related to the terrestrial atmosphere. These experiments are, however, limited in weight and cannot be easily manipulated during the flight that lasts from around ten hours to ten days. The improvement compared to ground-based experiments is, however, important and balloons have led the observational advances in the pre-WMAP era (BOOMERang [44], MAXIMA [45] and Archeops [46], for instance).
3. Satellite experiments allow a full sky survey to be performed but this solution is expensive. Only two such experiments have been conducted (COBE [47] and WMAP [48]) and a third one, Planck [49], is scheduled to be launched in 2009.

The revolution in this area has been led by the WMAP satellite, which has measured the cosmic microwave background anisotropies in five frequency bands (22, 30, 40, 60 and 90 GHz). Among the numerous results [50] of the three years of observation, we can note:

- the realization of a map of the temperature anisotropies with a resolution of 0.3 deg (Fig. 6.12) that gives access to the angular power spectrum up to $\ell \sim 1000$.
- a confirmation of the existence of the first acoustic peak, the position of which is now measured with a precision [51]

$$\ell_{(1)} = 220.8 \pm 0.7, \quad (6.254)$$

with an amplitude of $74.7 \pm 0.5 \mu\text{K}$. The second peak is located at $\ell_{(2)} = 530.9 \pm 3.8$ and has an amplitude of $48.8 \pm 0.9 \mu\text{K}$. The first and second troughs are located, respectively, at $\ell_{(-1)} = 412.4 \pm 1.9$ and $\ell_{(-2)} = 675.2 \pm 11.1$.

- a measure of the cross-correlation spectrum between the temperature and the E -polarization (Fig. 6.11). The spectrum has an antipeak at $\ell = 137 \pm 9$ and a peak at $\ell = 329 \pm 19$. The existence of this anticorrelation provides a new, even

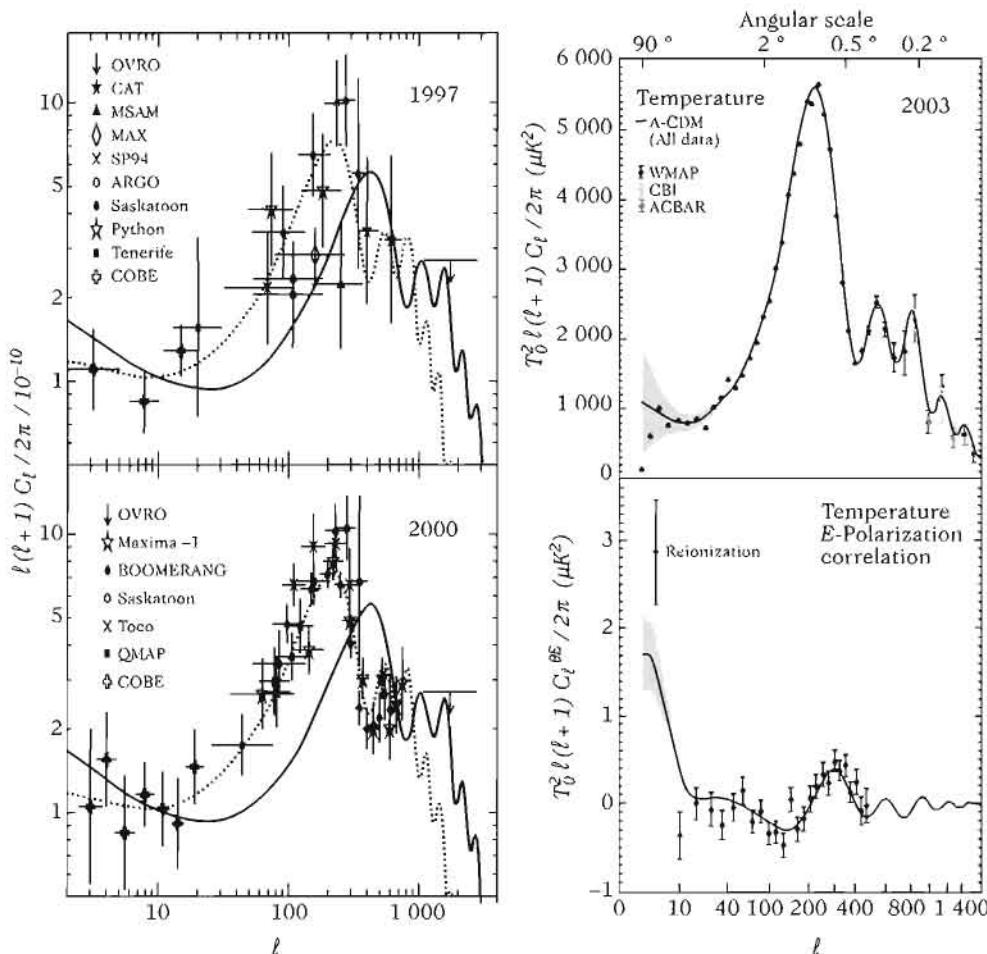


Fig. 6.11 Evolution of the observational status between 1997 and 2003 (left). The solid line curve represents the best-fit Λ CDM model, whereas an open model ($\Omega_0 = 0.3$, dotted line) is now excluded by observations. (right): The power spectra for the temperature and the temperature- E polarization cross-correlation obtained by the WMAP satellite.

more robust, proof of the existence of super-Hubble fluctuations at the time of decoupling, since the Sachs-Wolfe plateau of the temperature power spectrum alone could have been generated along the line of sight after decoupling (as, e.g., in topological defect models). The sign of the cross-correlation and the phase of the acoustic oscillations compared to that of the temperature power spectrum are strong indications in favour of adiabatic primordial perturbations.

- the ΘE cross-correlation at large scales ($\ell < 20$) is explained by an early reionization of the Universe, at a redshift of $z_{\text{re}} = 11.0^{+2.6}_{-2.5}$, which gives an optical depth of $\tau_{\text{re}} = 0.089 \pm 0.03$.

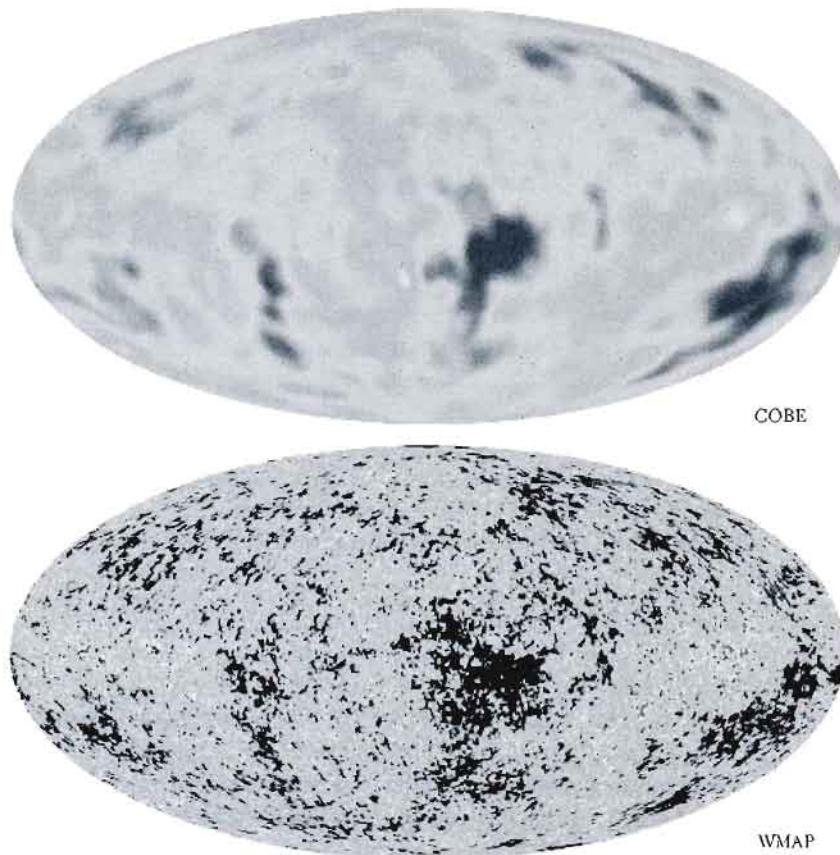


Fig. 6.12 Map of the cosmic microwave background anisotropies provided by WMAP compared to the resolution of the COBE satellite. (WMAP collaboration, 2003).

- The EE correlation was detected at a level of

$$\frac{\ell(\ell+1)}{2\pi} C_\ell^{EE} = (0.086 \pm 0.029)(\mu\text{K})^2$$

for $\ell = 2 - 6$. This is interpreted as the rescattering of cosmic microwave background photons at the reionization.

- No evidence for B -modes was seen, hence limiting

$$\frac{\ell(\ell+1)}{2\pi} C_\ell^{BB} = (-0.04 \pm 0.03)(\mu\text{K})^2$$

for $\ell = 2 - 6$.

- From these observations, it can be deduced that the temperature field is almost Gaussian, as predicted by inflation, but this question requires special attention.

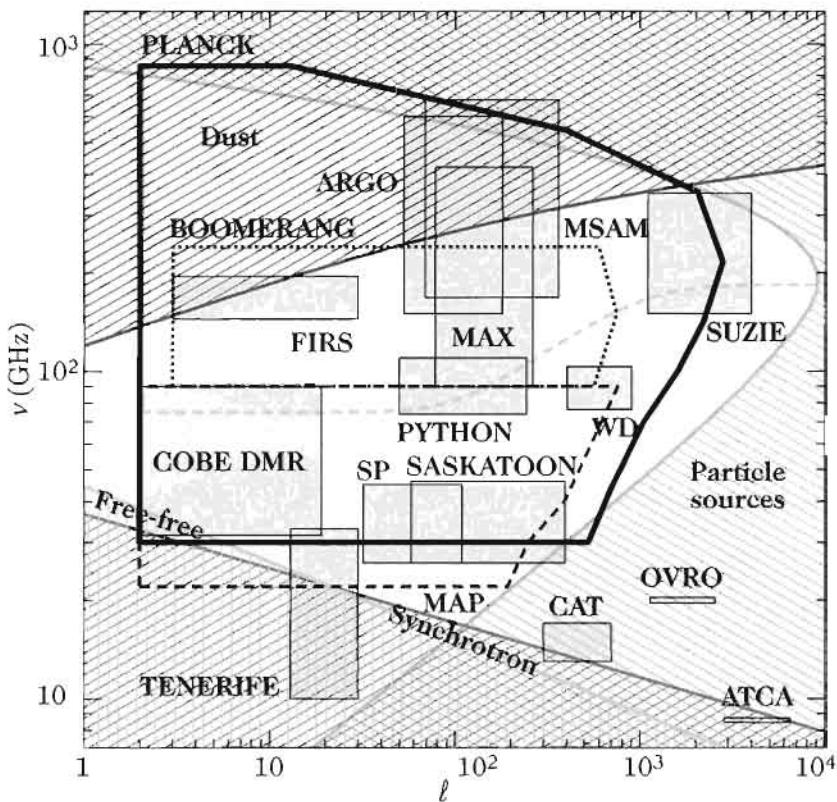


Fig. 6.13 Summary of the angular and frequency sensitivities of various cosmic microwave background observation experiments. The hatched zones correspond to the regions where the foreground emissions exceed $20 \mu\text{K}$ in the cleanest 20% region of the sky. From Ref. [53].

After WMAP, cosmology has a new standard model, the *best fit model* (or *concordance model*), which is a flat ΛCDM model with $\Omega_{\Lambda 0} \sim 0.7$. Even though observations are compatible with the predictions of inflation, some obscure points persist:

- WMAP has confirmed the fact that the quadrupole had less power than expected, as already pointed out by COBE.
- The octopole seems to be correlated with the quadrupole,
- Various statistical analyses seem to indicate differences between the northern and southern hemispheres. The existence and origin of this anisotropy is still not completely established.

The first detection of the polarization of the cosmic microwave background was announced in September 2002 [52]. This detection was obtained thanks to an interferometer, DASI, working at 30 GHz at the south pole. The amplitude of the *E* and *B* modes has been constrained by assuming that their spectrum was that of the best-fit ΛCDM model, which made it possible to establish a 5σ detection of the *E* modes.

The Planck satellite will observe the cosmic microwave background in nine frequency bands spread between 20 and 800 GHz with a resolution of the order of 4 arcmin, which corresponds to $\ell \sim 2000$. It uses radiometers, supposed to be 1000 times more sensitive than that of COBE, for the frequencies lower than 100 GHz and bolometers at high frequencies.

6.4.0.7 Statistical isotropy and observation on an incomplete sky

When observing the cosmic microwave background, various foreground emissions must be subtracted. This is particularly the case for the emission of our Galaxy, which in the simplest case, amounts to cutting out part of the observed sky. The observations then give access to

$$\tilde{\Theta}(e) = \Theta(e)W(e), \quad (6.255)$$

where W is a mask indicating which part of the sky has been cut out. Expanding the mask in spherical harmonics as

$$W(e) = \sum w_{\ell m} Y_{\ell m}(e), \quad (6.256)$$

with $w_{\ell m}^* = (-1)^m w_{\ell -m}$, then the coefficients $\tilde{a}_{\ell m}$ are related to the primordial coefficients, $a_{\ell m}$, by

$$\tilde{a}_{\ell m} = \sum_{\ell_1 m_1} a_{\ell_1 m_1} \sum_{\ell_2 m_2} w_{\ell_2 m_2} \int d^2 e Y_{\ell_1 m_1}(e) Y_{\ell_2 m_2}(e) Y_{\ell m}(e). \quad (6.257)$$

Using (B.23), this expression reduces to

$$\tilde{a}_{\ell m} = \sum_{\ell_1 m_1} a_{\ell_1 m_1} K_{\ell m}^{\ell_1 m_1}, \quad (6.258)$$

where the kernel $K_{\ell m}^{\ell_1 m_1}$ is explicitly given by

$$K_{\ell m}^{\ell_1 m_1} = (-1)^m \sum_{\ell_2 m_2} w_{\ell_2 m_2} \sqrt{\frac{(2\ell_1 + 1)(2\ell_2 + 1)(2\ell + 1)}{4\pi}} \begin{pmatrix} \ell_1 & \ell_2 & \ell \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \ell_1 & \ell_2 & \ell \\ m_1 & m_2 & -m \end{pmatrix}, \quad (6.259)$$

For a constant mask, $W = w_{00} Y_{00}$, we find that $K_{\ell m}^{\ell_1 m_1} = w_{00} \delta_{\ell\ell_1} \delta_{mm_1} / \sqrt{4\pi}$. We can then check that $\tilde{a}_{\ell m}$ does not satisfy the property $\langle \tilde{a}_{\ell m} \tilde{a}_{\ell' m'}^* \rangle = \delta_{\ell\ell'} \delta_{mm'} C_\ell$ but

$$\langle \tilde{a}_{\ell m} \tilde{a}_{\ell' m'}^* \rangle = \sum_{\ell_1 m_1} C_{\ell_1} K_{\ell m}^{\ell_1 m_1} K_{\ell' m'}^{\ell_1 m_1 *},$$

since the statistical isotropy is broken by the mask. This illustrates the difficulty in constructing an estimator of C_ℓ for a partial sky. Moreover, this stresses the difficulty in testing in practice the fact that the $a_{\ell m}$ are independent and isotropic stochastic fields.

6.5 Effects of the parameters on the angular power spectrum

The scalar, vector and tensor perturbations leave different signatures on the cosmic microwave background temperature anisotropies and polarizations. The *E*-polarization is generated by both the scalar and tensor modes, whereas the *B*-polarization is only induced by the tensor modes so that, in principle, it allows us to have a direct access to the properties of the gravitational waves. We can measure the power spectra and cross-correlations of these various quantities. Due to symmetry properties, only the correlation between Θ and E are non-vanishing. We thus have access to four observables ($\Theta\text{-}\Theta$, $E\text{-}E$, $\Theta\text{-}E$ and $B\text{-}B$); the first three are generated by the scalar and tensor modes while the fourth is uniquely generated by the tensor modes.

The theoretical curves of these angular power spectra depend on the value of the cosmological parameters and on the initial conditions for the perturbations. We illustrate here the influence of various parameters on the temperature power spectrum.

We shall distinguish the cosmological parameters that describe the matter content of the Universe and that affect the perturbations transfer function and the primordial parameters that describe the perturbations' initial conditions.

6.5.1 Cosmological parameters

6.5.1.1 Baryon density: $\Omega_{b0}h^2$

As seen earlier (see Section 6.2.2), the main parameter that influences the peak structure is R . Its value at the time of decoupling is given by

$$R_{\text{ess}} \sim 27.37 \Omega_{b0} h^2.$$

The baryon density Ω_{b0} has four main effects on the angular power spectrum.

- It fixes the oscillation frequency of the photon-baryon plasma via the sound speed. When Ω_{b0} increases, c_s decreases and the frequency of the oscillations is smaller. So, Ω_{b0} influences the position of the peaks and their spacing.
- It fixes the relative amplitude of the even and odd peaks. This amplitude is $1+6R$ in the adiabatic case. The greater Ω_{b0} , the smaller the second peak.
- It fixes the contrast between the peaks, by influencing the amplitude of the Doppler term. If $\Omega_{b0} = 0$, the peaks disappear altogether and the contrast increases when Ω_{b0} increases.
- It influences the Silk damping at small scales by changing the photons mean free path. The greater Ω_{b0} is, the shorter the mean free path and the more important is the damping.

Figure 6.14 illustrates these effects.

6.5.1.2 Total matter density: $\Omega_{m0}h^2$

In the standard model, the baryon density is small compared to that of dark matter, so that the total matter density can be approximated by that of dark matter. Unlike baryonic matter, dark matter only acts gravitationally on the photons. Ω_{m0} is thus only felt via its influence on the time evolution of the Bardeen potentials. In particular,

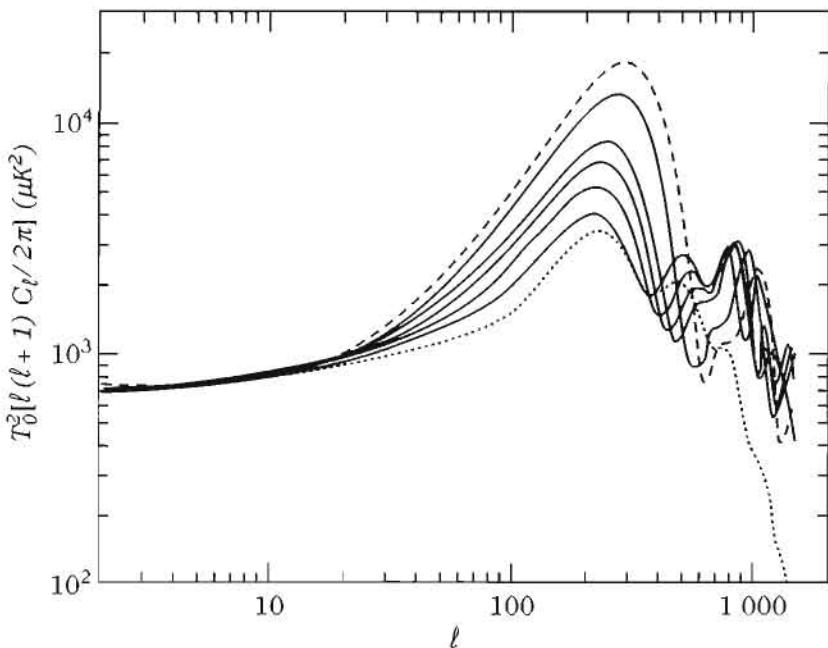


Fig. 6.14 Influence of the baryon density on the temperature angular power spectrum. For the spectra in solid lines, $\Omega_{b0} h^2 = 0.0125, 0.0250, 0.0375, 0.0500$ and 0.0875 . The two extremal values are 0.0025 (dotted) and 0.1250 (dashed). From Ref. [3].

an increase in Ω_{m0} is equivalent to an increase in z_{eq} and thus a delay in the matter-radiation equality, which affects the contribution of the early integrated Sachs-Wolfe term.

6.5.1.3 Dark-matter–baryon ratio

The baryon to dark matter ratio $\alpha = \Omega_b/\Omega_c$ has an influence on the relative contribution of both components in the Poisson equation. Since the density contrast of baryonic matter grows only after decoupling, we expect that the angular power spectrum will have a lower amplitude when α increases. Moreover, matter perturbations oscillate when they are coupled to photons and perturbations keep a trace of these oscillations after decoupling. These oscillations are then imprinted in the matter power spectrum and are known as baryonic acoustic oscillations.

6.5.1.4 Hubble constant: h

The Hubble constant is not actually a free parameter since, from the Friedmann equations, it cannot be varied independently from the matter content of the Universe. Its influence on the angular power spectrum depends on parameters that are kept constant when varying its value (the Ω_i or the $\omega_i = \Omega_i h^2$). As an example, if the curvature

and the cosmological constant are fixed, $1 + z_{\text{eq}} = \Omega_{m0}/\Omega_{r0} \simeq 2.4 \times 10^4 \Omega_{m0} h^2$. If h decreases, the matter–radiation transition occurs later, which has the effect of increasing the amplitude of the acoustic peaks (Fig. 6.15).

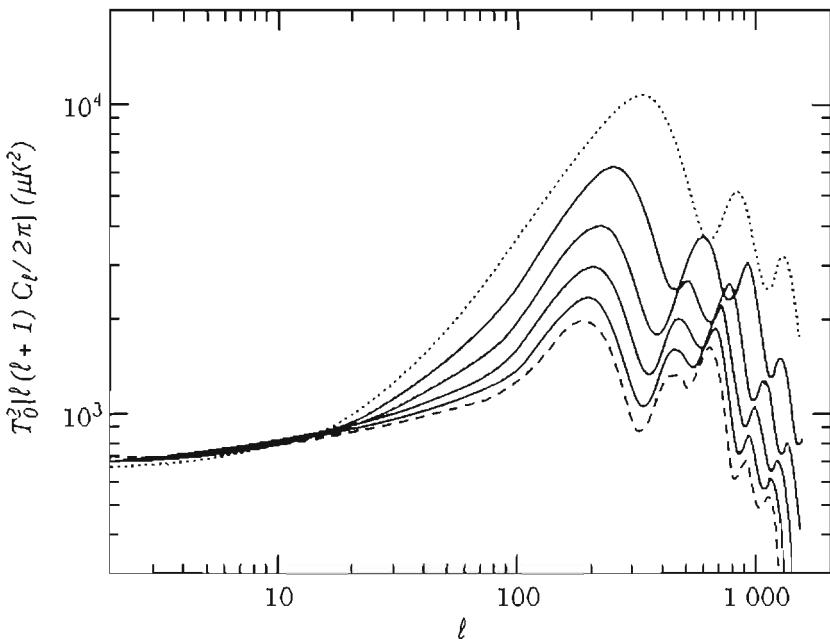


Fig. 6.15 Influence of the Hubble constant. The baryon density is fixed to $\Omega_{b0} = 0.0125$ and that of dark matter varies with h that takes the following values 0.2 (dotted line), 0.35, 0.50, 0.65, 0.80 (solid lines) and 0.95 (dashed lines). From Ref. [3].

6.5.1.5 Cosmological constant: $\Omega_{\Lambda 0}$

Varying $\Omega_{\Lambda 0}$ at fixed Ω_0 , the cosmological constant delays the matter–radiation equality and modifies the relation between the angular distance and the redshift. The position of the acoustic peaks will thus be affected (see Fig. 6.16).

Moreover, the cosmological constant dominates today so that the Bardeen potentials decay at late time, which induces a late integrated Sachs–Wolfe effect and increases the power at large angular scales.

6.5.1.6 Curvature: $\Omega_{K0} h^2$

The effect of the curvature is similar to that of the cosmological constant. In an open Universe, the relation between the angular distance and the redshift is modified and the Bardeen potentials decay after the matter–curvature equality. The effect of the angular distance implies also that the acoustic peaks will be shifted towards higher multipoles for open spaces and smaller ones for closed spaces.

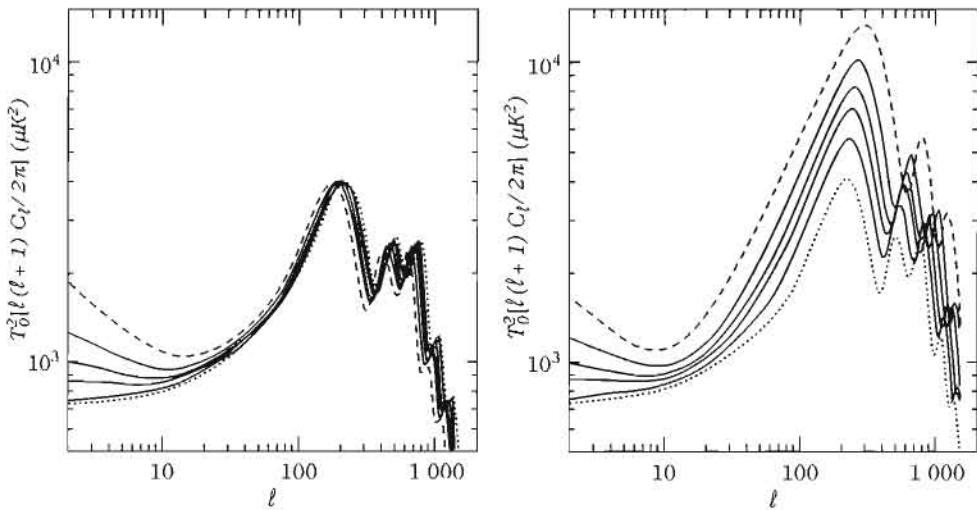


Fig. 6.16 Influence of the cosmological constant on the temperature angular power spectrum. (left): $\Omega_{\Lambda 0}$ varies while keeping $\rho_{crit 0}$ and $\bar{\rho}_0$ fixed, which implies that h also varies. $\Omega_{\Lambda 0}$ takes the values 0 (dotted line), 0.4, 0.6, 0.7, 0.8 (solid lines) and 0.9 (dashed line). (right): $\rho_{b 0}$ and h are kept constant so that $\Omega_{m 0}$ varies as $1 - \Omega_{\Lambda 0}$. $\Omega_{\Lambda 0}$ takes the values 0 (dashed line), 0.4, 0.6, 0.7, 0.8 (solid lines) and 0.9 (dotted line). From Ref. [3].

Notice that we can simultaneously adjust the curvature and the cosmological constant while keeping the spectrum almost unchanged. These effects are illustrated in Fig. 6.17.

6.5.1.7 Parameter degeneracy

As illustrated in the previous figures, varying the cosmological parameters affects, on the one hand, the position of the peaks and their relative amplitude and, on the other, the shape of the Sachs–Wolfe plateau. At large angular scales, the cosmic variance becomes more important and different sets of parameters can lead to identical angular power spectra up to the cosmic variance precision.

Actually, these degeneracies can be understood if we remember that the position of the peaks is mainly dictated by the acoustic scale $\ell_A = f_K(z_{LSS})/r_s(z_{LSS})$ (6.74). For instance, the sensitivity of this scale to the standard cosmological parameters is

$$\frac{\Delta \ell_A}{\ell_A} = -1.1 \frac{\Delta \Omega_0}{\Omega_0} - 0.24 \frac{\Delta \Omega_{m0} h^2}{\Omega_{m0} h^2} - 0.17 \frac{\Delta \Omega_{\Lambda 0}}{\Omega_{\Lambda 0}} - 0.11 \Delta w_Q - 0.07 \frac{\Delta \Omega_{b0} h^2}{\Omega_{b0} h^2},$$

around the model $(\Omega_0, \Omega_{m0} h^2, \Omega_{\Lambda 0}, w_Q, \Omega_{b0} h^2) = (1, 0.15, 0.65, -1, 0.02)$, w_Q being the equation of state of dark matter. These degeneracies are illustrated in Fig. 6.18.

6.5.2 Parameters describing the primordial physics

Primordial physics provides the initial power spectra for scalar and tensor perturbations that are characterized by their amplitudes, spectral indices and variation of

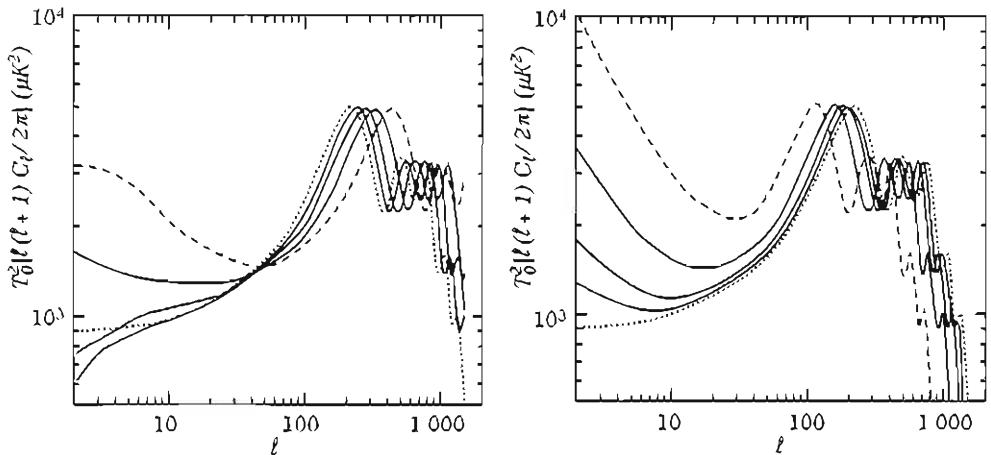


Fig. 6.17 Influence of the curvature on the angular power spectrum. (left): Spatially hyperbolic spaces. The matter density is constant and Ω_0 is, respectively, 1 (dotted line), 0.8, 0.6, 0.4 (solid lines) and 0.2 (dashed line). The peaks are shifted towards smaller angular scales and the decay of the gravitational potentials induces a late integrated Sachs–Wolfe effect with opposite sign to that of the ordinary Sachs–Wolfe term. (right): Spatially spherical spaces. The matter density is constant and Ω_0 is, respectively, 1 (dotted line), 1.2, 1.4, 2.0 (solid lines) and 4.0 (dashed line). The Bardeen potentials increase in time since Ω_m increases before the expansion stops. The integrated Sachs–Wolfe term has thus the same sign as the Sachs–Wolfe term, which translates into an increase in the large-scale power. From Ref. [3].

their spectral indices (Chapter 8). It also provides an indication on the adiabatic or isocurvature nature of the initial conditions for the perturbations.

6.5.2.1 Spectral index

As can be seen from (6.50) and (6.96), the quantity $\ell(\ell+1)C_\ell$ is proportional to ℓ^{n_s-1} and ℓ^{n_T} for the scalar and tensor modes, respectively.

For a red-tilted spectrum ($n_s < 1$, $n_T < 0$) the angular power spectrum has a power excess at large angular scales, while a blue-tilted spectrum has a power excess at short angular scales.

A deviation primordial spectrum from a scale invariance thus induces a tilt in the Sachs–Wolfe plateau and a modification of the relative amplitude of the acoustic peaks compared with the Sachs–Wolfe plateau. This effect is illustrated in Fig. 6.19.

6.5.2.2 Primordial gravitational waves

The contribution of the gravitational waves is dominant on large angular scales, so that they tend to decrease the relative height of the acoustic peaks (of purely scalar origin) compared to the Sachs–Wolfe plateau.

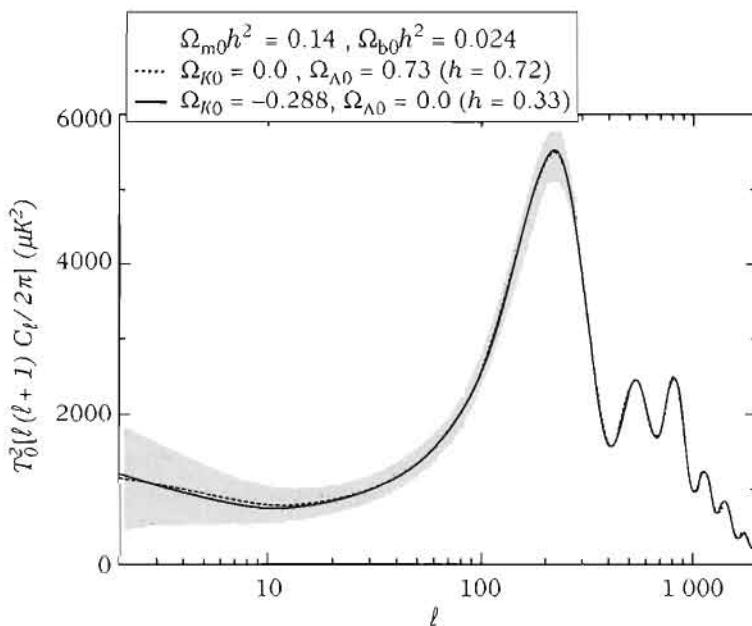


Fig. 6.18 Illustration of the degeneracy of the cosmological parameters. Both models are indistinguishable given the limitation due to the cosmic variance. They have in common $(\Omega_{m0}h^2, \Omega_{b0}h^2) = (0.14, 0.024)$. The first model (dotted line) is the best-fit Λ CDM model with $(\Omega_{K0}, \Omega_{\Lambda0}) = (0, 0.73)$ and $h = 0.72$, whereas the second (solid line) is an open model with a low Hubble constant, $(\Omega_{K0}, \Omega_{\Lambda0}) = (-0.288, 0)$ and $h = 0.33$.

The structure of the acoustic peaks is still that of the scalar modes and the normalization of the scalar modes spectrum is lower when the gravitational waves contribute at large angular scales.

6.5.2.3 Adiabatic and isocurvature initial conditions

The choice in the type of initial conditions influences the peaks position and their relative amplitude. This is due to the relation (6.74) that explains how the peaks structure is modified. In particular, the position of the first peak is shifted by a factor 3/2 for an isocurvature model.

A second difference arises in the normalization of the spectra since $\Theta_{SW} = \Phi/3$ for a pure adiabatic model and 2Φ for a pure isocurvature model.

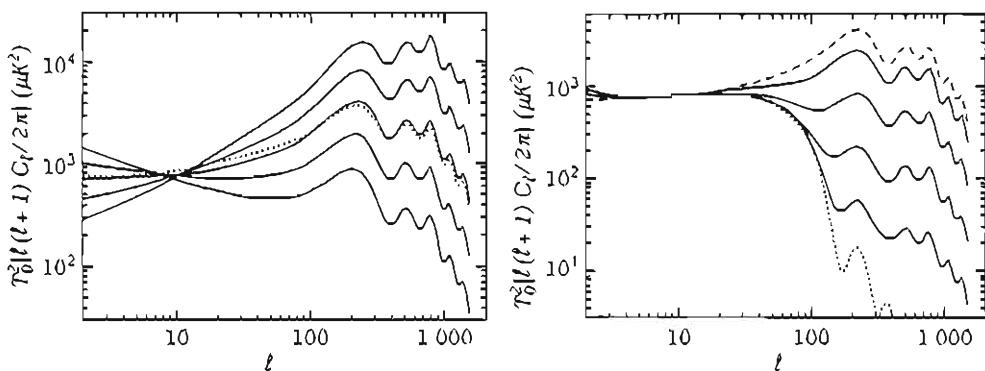


Fig. 6.19 (left): Influence of the spectral index of the scalar modes on the primordial spectrum. n_s takes the values 0.5, 0.75, 1.0, 1.25 and 1.5 (solid lines) from bottom to top in the Doppler region. The dotted spectrum corresponds to the case $n_s = 1.5$ multiplied by $l^{-0.5}$, in agreement with the spectrum obtained for $n_s = 1$. (right): Influence of the T/S ratio. The dotted and dashed spectra represent the contributions from the scalar and tensor modes alone, and we interpolate between these two curves when T/S varies from 0 to ∞ . From Ref. [3].

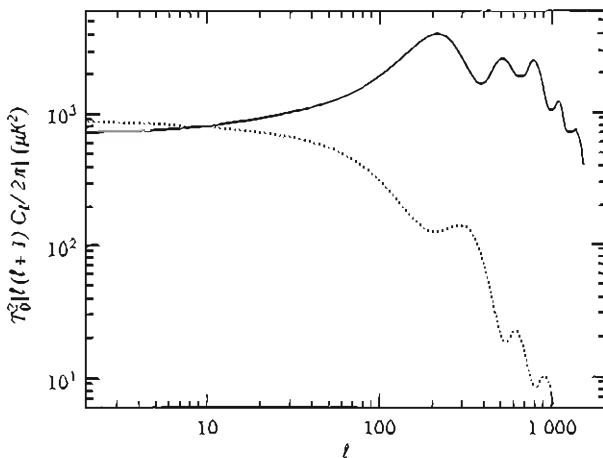


Fig. 6.20 Comparison of the temperature angular power spectrum for adiabatic and isocurvature initial conditions. From Ref. [3].

References

- [1] R. DURRER, ‘The theory of CMB anisotropies’, *J. Phys. Stud.* **5**, 177, 2001.
- [2] W. HU, *Wandering the background: a cosmic background explorer*, Thesis, University of Berkeley, 1995, [[astro-ph/9508126](#)].
- [3] A. RIAZUELO, *Signature de divers modèle d'univers primordial dans les anisotropies du rayonnement fossile*, PhD thesis, University of Paris XI, 2000
- [4] A. KOSOWSKY, *The cosmic microwave background*, in *Modern cosmology*, S. Bonometto et al. (eds.), Institute of Physics, 2002.
- [5] R.K. SACHS and A.M. WOLFE, ‘Perturbations of a cosmological model and angular variations of the microwave background’, *Astrophys. J.* **147**, 73, 1967.
- [6] M. PANEK, ‘Large-scale microwave background fluctuations: gauge invariant formalism’, *Phys. Rev. D* **34**, 416, 1986.
- [7] W. HU and N. SUGIYAMA, ‘Toward understanding CMB anisotropies and their implications’, *Phys. Rev. D* **51**, 2599, 1995; W. HU and N. SUGIYAMA, ‘Anisotropies in the CMB: an analytic approach’, *Astrophys. J.* **444**, 489, 1995.
- [8] A. SAKHAROV, *Zh. Eksp. Teor. Fiz.* **49**, 345, 1965, [*Sov. Phys. JETP* **22**, 241, 1966].
- [9] J. SILK, ‘Cosmic black-body radiation and galaxy formation’, *Astrophys. J.* **151**, 459, 1968.
- [10] M. BIRKINSHAW, *The Sunyaev-Zeldovich effect*, *Phys. Rep.* **310** (1999) 97.
- [11] J. BERNSTEIN, *Kinetic theory in the expanding Universe*, Cambridge University Press, 1988.
- [12] J. STEWART, *Non-equilibrium relativistic kinetic theory*, *Lect. Notes in Phys.* **10**, Springer, Berlin, 1971.
- [13] J.-P. UZAN, ‘Dynamics of relativistic interacting gases: from a kinetic to a fluid description’, *Class. Quant. Grav.* **15**, 1063, 1998.
- [14] M. HILLERY, ‘Distribution functions in physics: fundamentals’, *Phys. Rep.* **106**, 121, 1984.
- [15] P.J.E. PEEBLES and T.J. YU, ‘Primeval adiabatic perturbation in an expanding Universe’, *Astrophys. J.* **162**, 815, 1970.
- [16] J.R. BOND and G. EFSTATHIOU, ‘Cosmic background radiation anisotropies in Universes dominated by nonbaryonic dark matter’, *Astrophys. J. Lett.* **285**, L45, 1984.
- [17] C.P. MA and E. BERTSCHINGER, ‘Cosmological perturbation theory in the synchronous and conformal gauges’, *Astrophys. J.* **455**, 7, 1995.
- [18] H. KODAMA and M. SASAKI, ‘Cosmological perturbation theory’, *Prog. Theor. Phys. Suppl.* **78**, 1, 1984.
- [19] C. PITROU, ‘Gauge invariant Boltzmann equation and the fluid limit’, *Class. Quant. Grav.* **24**, 6127, 2007.

- [20] K. THORNE, 'Multipole expansions of gravitational radiation', *Rev. Mod. Phys.* **52**, 299, 1980.
- [21] R. DURRER, 'Gauge invariant cosmological perturbation theory: a general study and its application to the texture scenario of structure formation', *Fund. Cosm. Phys.* **14**, 209, 1994.
- [22] M.J. REES, 'Polarization and spectrum of the primeval radiation in an anisotropic Universe', *Astrophys. J.* **153**, L1, 1968.
- [23] W. HU and M. WHITE, 'A CMB polarization primer', *New Astron.* **2**, 323, 1997.
- [24] M. ZALDARRIAGA and D. HARARI, 'Analytic approach to the polarization of the cosmic microwave background in flat and open Universes', *Phys. Rev. D* **52**, 3276, 1995.
- [25] A. KOSOWSKY, 'Cosmic microwave background polarization', *Annals. Phys.* **246**, 49, 1995.
- [26] M. ZALDARRIAGA, *The polarization of the cosmic microwave background*, in *Measuring and Modeling the Universe*, Ed. W. L. Freedman, Carnegie Observatories Astrophysics Series, 309, CUP 2004 [[astro-ph/0305272](#)].
- [27] W. HU and M. WHITE, 'CMB anisotropies: total angular momentum method', *Phys. Rev. D* **56**, 596, 1997.
- [28] J. JACKSON, *Classical electrodynamics*, Wiley, 1998.
- [29] E. NEWMAN and R. PENROSE, *J. Math. Phys.* **7**, 863, 1966; N.J. VILENKIN, *Special functions and the theory of group representations*, American Mathematical Society Press, Providence, 1968.
- [30] S. CHANDRASEKHAR, *Radiative transfer*, Dover, 1960.
- [31] M. KAMIONKOWSKI et al., 'A probe of primordial gravity waves and vorticity', *Phys. Rev. Lett.* **78**, 2058, 1997; M. KAMIONKOWSKI et al., 'Statistics of cosmic microwave background polarization', *Phys. Rev. D* **55**, 7368, 1997; W. HU et al., 'A complete treatment of CMB anisotropies in a FRW Universe', *Phys. Rev. D* **57**, 3290, 1998.
- [32] U. SELJAK, 'A two-fluid approximation for calculating the cosmic microwave background anisotropies', *Astrophys. J.* **435**, L87, 1994.
- [33] U. SELJAK and M. ZALDARRIAGA, 'A line of sight integration approach to cosmic microwave background anisotropies', *Astrophys. J.* **469**, 437, 1996.
- [34] CMBFAST: <http://cmbfast.org>, developed by U. Seljak and M. Zaldarriaga.
- [35] CAMB: <http://camb.info>, developed by A. Lewis and A. Challinor.
- [36] CMBEASY: <http://www.thphys.uni-heidelberg.de/~doran/cmbeasy>, developed by C. Doran.
- [37] COSMICS: <http://acturus.mit.edu/cosmics>, developed by E. Bertschinger and C.P. Ma.
- [38] G.F. SMOOT, 'CMB experiments', *Phys. Rep.* **333**, 269, 2000.
- [39] Report 'Le rayonnement fossile à 3 K', *C.R. Acad. Sci. (Paris)* **4** (2003); S. EIDELMAN et al., Particle Data Group, Chap. 23 'Cosmic microwave background', *Phys. Lett. B* **592**, 1, 2004.
- [40] A.W. JONES and A.N. LASENBY, 'The cosmic microwave background', *Living Rev. Relativity* (1998)
- [41] DASI: <http://astro.uchicago.edu/dasi>.

- [42] CAT: <http://www.mnrao.cam.ac.uk/telescopes/cat/index.html>.
- [43] VLA: <http://www.nrao.edu/vla/html/VLAhome.shtml>.
- [44] BOOMERanG: <http://www.physics.ucsb.edu/~boomerang/>.
- [45] MAXIMA: <http://cfpa.berkeley.edu/group/cmb/>.
- [46] ARCHEOPS: <http://www.archeops.org>.
- [47] COBE:
http://space.gsfc.nasa.gov/astro/cobe/cobe_home.html.
- [48] WMAP: <http://map.gsfc.nasa.gov>
- [49] Planck Surveyor: <http://astro.estec.esa.nl/Planck>.
- [50] L. PAGE *et al.*, 'Three year WMAP observations: Polarization analysis', *Astrophys. J. Suppl.* **170**, 335, 2007.
- [51] D.N. SPERGEL *et al.*, 'WMAP three year results: implication for cosmology', *Astrophys. J. Suppl.* **170**, 377, 2007.
- [52] J.M. KOVAC *et al.*, 'Detection of polarization in the cosmic microwave background using DASI', *Nature* **420**, 772, 2002.
- [53] M. TEGMARK CMB analysis centre:
<http://space.mit.edu/home/tegmark/cmb/experiments.html>.

7

Gravitational lensing and dark matter

7.1 Gravitational lensing and its applications

Light is deflected by any inhomogeneous gravitational field (see Chapter 1). This property was at the origin of the first tests of general relativity in the Solar System. Today, it has become a powerful tool for mapping the distribution of matter in the Universe, and in particular that of dark matter. The applications of gravitational lensing are numerous. Table 7.1 gives a classification of the various gravitational lensing systems and their physical applications.

Various complementary studies can be found in the two monographs [1, 2] and in the recent reviews [3, 4] that detail the historical, theoretical and observational aspects.

Table 7.1 Classification of the different (most studied, with well-established observations) gravitational lensing systems in terms of the source and the lens. The effects are classified into strong lensing (SL), microlensing (ML), weak lensing (WL) and statistical (st).

| Source | Lens | Effects | Applications |
|----------------------------|------------------------------|--|---|
| Star (galactic) | Star (galactic) | ML: light curves | Extrasolar planets Milky Way inner structure |
| Star (extragalactic) | Compact object (galactic) | ML: light curves | Dark matter in Milky Way halo |
| Galaxy | Galaxy | stWL: ellipticity bias, shear-galaxy correlation | Galaxy parameters, halos properties |
| Galaxy | Cluster | SL: giant arcs, multiple image WL: arclets, magnification bias | Total mass of the cluster, Ω_Λ , redshifts Cluster mass profile, cluster morphology |
| Galaxy | Large-scale structure | StWL: cosmic shear, shear-shear correlation | Cosmological parameters Cosmological power spectrum, |
| Quasar | Galaxy | SL: multiple images, Einstein rings, | Mass of galaxies, substructures, H_0 , Ω_Λ |
| Quasar | Large-scale structure | StWL: cosmic magnification, quasar-galaxy correlation | cosmological parameters, power spectrum, bias |
| Last scattering surface | Large scale structure | StWL: smoothing of C_ℓ , B modes | Cosmological parameters, power spectrum |

7.1.1 Gravitational lensing in the thin-lens regime

7.1.1.1 Light deflection by a point-like mass

Lens equation

The lens equation relates the real position of a source to its observed position (see Fig. 7.1). First, we define the *optical axis* as the line joining the observer to the centre of the lens (OL). The *source plane* and the *lens plane* are the two planes perpendicular to the optical axis located at distances D_{OS} and D_{OL} , respectively, from the observer.

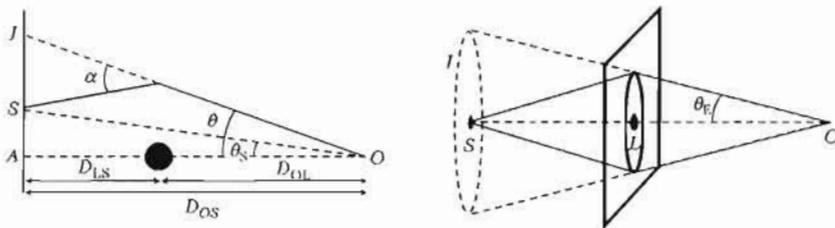


Fig. 7.1 (left): Gravitational lens configuration. L is the gravitational lens. θ , θ_S and α are, respectively, the angular position of the image, the angular position of the source and the deflection angle. D_{OL} and D_{OS} are the angular distances between the observer and the lens and between the observer and the source, respectively, and D_{LS} is the angular distance between the lens and the source. (right): When the source, the lens and the observer are aligned, the image of the source takes the shape of a circle, called an Einstein ring.

The positions of the source and of its image on the source plane are related by $AI = AS + SI$, where AS can be expressed in terms of the angular position of the source θ_S by $AS = \theta_S D_{OS}$ and similarly, AI is given by $AI = \theta D_{OS}$, with θ the angular position of the image, and $SI = \alpha D_{LS}$, α being the deflection angle. We thus find that in the small-angle limit, the angular position of the image $\theta = \theta_L$ is

$$\theta = \theta_S + \frac{D_{LS}}{D_{OS}} \alpha(\theta). \quad (7.1)$$

Notice that the distances involved in this equation are the angular distances since they relate transverse physical distances to their angular diameter. The value of the deflection angle by a point-like source was obtained in Chapter 1 [see (1.143)]; $\alpha = 4G_N M/bc^2 = 2R_{\text{Sch}}/b$, where $b = D_{OL}\theta$ is the impact parameter and R_{Sch} is the Schwarzschild radius. The lens equation therefore becomes

$$\theta = \theta_S + 2 \frac{R_{\text{Sch}}}{\mathcal{D}} \frac{1}{\theta}, \quad (7.2)$$

where we have introduced the geometrical factor $\mathcal{D} \equiv D_{OS}D_{OL}/D_{LS}$.

Einstein radius

The rotational invariance around the optical axis implies that a source located exactly on this axis ($\theta_S = 0$) will be observed as a ring of angular diameter

$$\theta_E = \sqrt{\frac{4G_N M}{c^2} \frac{D_{LS}}{D_{OS} D_{OL}}}. \quad (7.3)$$

This angular radius, called the *Einstein radius*, depends both on the lens mass and on the factor $D_{LS}/D_{OS}D_{OL}$ that characterizes the geometry of the system. Its physical radius is $R_E = \theta_E D_{OL}$ (see Fig. 7.1).

The Einstein radius is a natural angular scale for the problem. If lensing gives rise to multiple images, $2\theta_E$ will be the typical angular separation between these images. Sources closer than θ_E from the optical axis are subject to strong distortion effects, giving rise, for instance, to arcs and will be strongly amplified and distorted. Sources with $\theta_S \gg \theta_E$ are almost unaffected by the lens. The typical orders of magnitude of the Einstein radius for gravitational lensing by a star in the Galaxy ($M \sim M_\odot$ and $D \sim 10$ kpc) and by a galaxy ($M \sim 10^{12} M_\odot$) in the cosmological context ($D \sim D_{H_0}$) are

$$\theta_{E,\text{star}} = 0.9 \left(\frac{M}{M_\odot} \right)^{1/2} \left(\frac{D}{10 \text{ kpc}} \right)^{-1/2} \text{ mas}, \quad (7.4)$$

$$\theta_{E,\text{gal}} = 2 \left(\frac{M}{10^{12} M_\odot} \right)^{1/2} \left(\frac{D}{3000 \text{ Mpc}} \right)^{-1/2} \text{ arcsec}. \quad (7.5)$$

For instance, for a lens at $z = 0.5$ and a source at $z = 2$, we obtain $\theta_E \sim 2.13$ arcsec for an Einstein-de Sitter Universe with $h = 0.7$.

Multiple images and magnification

Equation (7.2) can be rewritten in terms of the Einstein radius in the simplified form

$$\theta_S = \theta - \frac{\theta_E^2}{\theta}. \quad (7.6)$$

Solving this equation for θ gives two solutions

$$\theta_{\pm} = \frac{1}{2} \left(\theta_S \pm \sqrt{\theta_S^2 + 4\theta_E^2} \right), \quad (7.7)$$

so that each source has two gravitational images. These two images are located on either side of the source, one inside the Einstein radius ($\theta_- < \theta_E$) and the other outside ($\theta_+ > \theta_E$), and are separated by $\Delta\theta \geq 2\theta_E$.

Since the light deflection is not associated with any absorption or emission, the gravitational lensing conserves the surface brightness. The surface brightness of the image, $I(\theta)$, is related to that of the source, $I_S(\theta_S)$, by

$$I(\theta) = I_S |\theta_S(\theta)|. \quad (7.8)$$

The lens gravitational field distorts the light beam and modifies the solid angle under which the source is observed. The total flux is thus proportional to the ratio between

the solid angles under which the source and the image are observed. For a symmetric lens, such as a point-like lens, the magnification is given directly by

$$\mu = \frac{\theta}{\theta_S} \frac{d\theta}{d\theta_S}, \quad (7.9)$$

since the surface element is proportional to $\sin \theta d\theta \sim \theta d\theta$, for small angles. In the case of a point-like lens, we obtain, after using (7.6),

$$\mu_{\pm} = \left[1 - \left(\frac{\theta_E}{\theta_{\pm}} \right)^4 \right]^{-1} = \frac{1}{2} \pm \frac{u^2 + 2}{2u\sqrt{u^2 + 4}}, \quad (7.10)$$

where $u \equiv \theta_S/\theta_E$ is the impact parameter in units of the Einstein radius. Since $\theta_- < \theta_E$, $u_- < 1$ and $\mu_- < 0$. This means that the image inside the Einstein radius has an inverted parity; it is the image of the source in a mirror. Figure 7.2 represents the position and the relative orientation of the two images.

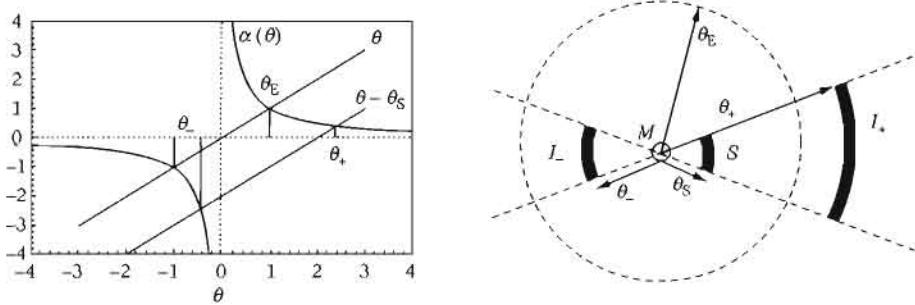


Fig. 7.2 (left): The solution of the lens equation can be obtained graphically through the intersection of the curve $\alpha(\theta)$ and the line of equation $\theta - \theta_S$. (right): The relative position of the two images of the source and their relative magnification.

The total magnification is defined as the sum of the magnification of each image.

$$\mu = |\mu_+| + |\mu_-| = \frac{u^2 + 2}{u\sqrt{u^2 + 4}}. \quad (7.11)$$

Moreover, one can check that $\mu_+ + \mu_- = 1$. Notice that μ is always greater than 1, for instance, for a source on the Einstein radius, $u = 1$ and $\mu \simeq 1.34$. As the source gets closer to the optical axis, θ_S tends toward zero and the magnification diverges, hence the magnification of an Einstein ring would be infinite. This divergence disappears for objects that are not point-like (and is thus an artefact of this approximation).

7.1.1.2 General case

Thin-lens approximation

The previous example illustrates the different effects produced by a gravitational lens, in the approximation in which it is point-like, an approximation beyond which we need to go. Since a real lens is a priori not symmetric under a rotation around the optical axis, we must replace the angles by two-dimensional vectors representing the angular position in the sky, assuming it is flat.

As seen in Chapter 1, the total deflection to which a photon is subjected is

$$\alpha = \frac{2}{c^2} \int \nabla_{\perp} \Phi dz, \quad (7.12)$$

where ∇_{\perp} represents a two-dimensional gradient perpendicular to the line of sight and Φ is the gravitational potential. In the case of a point-like lens of mass M , this potential is explicitly given by $\Phi = -G_N M / \sqrt{b^2 + z^2}$ where b is the impact parameter of the undeflected light beam and z is the coordinate along the optical axis. This implies that $\nabla_{\perp} \Phi = G_N M b / (b^2 + z^2)^{3/2}$ and, after integration over z , that $\alpha = 4G_N M b / c^2 b^2$.

As long as the deflection angle is small, the light ray is only weakly deviated with respect to the undeflected trajectory. The integral (7.12) can thus be computed by integration along the undeflected trajectory. Such a Born approximation is completely justified for galaxies and galaxy clusters for which the deflection angles are typically $\alpha \lesssim 1''$ and $\alpha \lesssim 30''$, respectively.

Moreover, the main contribution to the integral (7.12) comes from the interval $\Delta z \sim \pm b$ around the lens. This distance over which the deflection is effective is very small compared to the other distances along the optical axis. We can therefore assume that the lens is thin in the direction of the line of sight and that its three-dimensional mass distribution, ρ_L , can be replaced by a two-dimensional mass distribution, Σ , obtained by integrating along the line of sight

$$\Sigma(b) \equiv \int \rho_L(b, z) dz. \quad (7.13)$$

This simplifying approximation is well adapted to the study of the lensing by galaxies or galaxy clusters. It encodes the entire information on the deflector in a single function, $\Sigma(b)$, its surface mass density, localized in the lens plane at D_{OL} . In the case of the point-like lens studied in Section 7.1.1.1 we have $\Sigma(b) = M \delta^{(2)}(b)$ and thus, in terms of angular coordinates, $\Sigma(\theta) = (M/D_{OL}^2) \delta^{(2)}(\theta)$.

Lens equation

In the approximation discussed above, the lens equation (7.1) takes the vector form

$$\theta = \theta_s + \frac{D_{LS}}{D_{OS}} \alpha(\theta). \quad (7.14)$$

To express α as a function of Σ , let us note that the three-dimensional Laplacian of the gravitational potential can be decomposed into $\Delta \Phi = \Delta_{\perp} \Phi + \partial^2 \Phi / \partial z^2$. The Poisson equation (5.5) can then be rewritten as

$$\Delta_{\perp} \Phi = 4\pi G_N \rho_L(\mathbf{b}, z) - \frac{\partial^2 \Phi}{\partial z^2}, \quad (7.15)$$

so that the deflection angle is the solution of

$$\nabla_{\perp} \cdot \boldsymbol{\alpha} = \frac{2}{c^2} \int_{-\infty}^{\infty} \left[4\pi G_N \rho_L(\mathbf{b}, z) - \frac{\partial^2 \Phi}{\partial z^2} \right] dz. \quad (7.16)$$

The integral of the term $(\partial^2 \Phi / \partial z^2)$ vanishes and after integrating over z , we obtain

$$\nabla_{\perp} \cdot \boldsymbol{\alpha} = \frac{8\pi G_N}{c^2} \Sigma(\mathbf{b}), \quad (7.17)$$

which is solved by

$$\boldsymbol{\alpha} = \frac{4G_N}{c^2} \int d^2 b' \Sigma(b') \frac{\mathbf{b} - \mathbf{b}'}{|\mathbf{b} - \mathbf{b}'|^2}. \quad (7.18)$$

In order to avoid any confusion, we will denote by Δ the usual three-dimensional Laplacian; Δ_{\perp} the two-dimensional Laplacian in the physical space (i.e. with respect to \mathbf{b}) and Δ_2 the two-dimensional Laplacian in angle space (i.e. with respect to $\boldsymbol{\theta}$) so that $\Delta_2 = D_{OL}^2 \Delta_{\perp}$. We shall use the same convention for the gradient. After introducing the critical surface density

$$\Sigma_{crit} \equiv \frac{c^2}{4\pi G_N} \frac{D_{OS}}{D_{OL} D_{LS}} \simeq 0.35 \left(\frac{\mathcal{D}}{1 \text{ Gpc}} \right)^{-1} \text{ g} \cdot \text{cm}^{-2}, \quad (7.19)$$

the lens equation in the thin-lens regime reduces to

$$\theta = \theta_S + \hat{\alpha}(\theta), \quad \hat{\alpha}(\theta) = \frac{1}{\pi} \int d^2 \theta' \frac{\Sigma(\theta')}{\Sigma_{crit}} \frac{\theta - \theta'}{|\theta - \theta'|^2}.$$

(7.20)

A lens with constant surface density equal to Σ_{crit} induces a deflection angle $\hat{\alpha} = \theta$ so that $\theta_S = 0$ for all θ . Such a lens perfectly focuses light and a focal length can be associated with it. This is not the case for arbitrary lenses that distort the images and are thus astigmatic. It is also useful to compare Σ_{crit} to the typical value of the surface density for a galaxy and a cluster, $\Sigma \sim M/R^2$,

$$\Sigma_{gal} \simeq 0.3 \left(\frac{M}{10^{11} M_{\odot}} \right) \left(\frac{R_{gal}}{10 \text{ kpc}} \right)^{-2} \text{ g} \cdot \text{cm}^{-2}, \quad (7.21)$$

$$\Sigma_{cluster} \simeq 0.03 \left(\frac{M}{10^{15} M_{\odot}} \right) \left(\frac{R_{cluster}}{3 \text{ Mpc}} \right)^{-2} \text{ g} \cdot \text{cm}^{-2}. \quad (7.22)$$

Projected potential

It is useful to introduce the projected gravitational potential obtained by integrating along the line of sight

$$\psi(\theta) = \frac{D_{\text{LS}}}{D_{\text{OL}} D_{\text{OS}}} \frac{2}{c^2} \int \Phi(D_{\text{OL}}\theta, z) dz. \quad (7.23)$$

Using $\nabla_2 \ln \theta = \theta/\theta^2$, the above potential is related to the deflection angle (7.12) through

$$\nabla_2 \psi(\theta) = \frac{D_{\text{LS}}}{D_{\text{OS}}} \alpha. \quad (7.24)$$

Using (7.17) and the property $\Delta_2 \ln \theta = 2\pi\delta^{(2)}(\theta)$, we finally obtain

$$\frac{1}{2} \Delta_2 \psi(\theta) = \frac{\Sigma}{\Sigma_{\text{crit}}}. \quad (7.25)$$

The solution of this two-dimensional Poisson equation gives us the expression for the projected potential as a function of the surface density

$$\psi(\theta) = \frac{1}{\pi} \int d^2\theta' \frac{\Sigma(\theta')}{\Sigma_{\text{crit}}} \ln |\theta - \theta'|. \quad (7.26)$$

Magnification and distortion

Equation (7.8) implies that

$$I(\theta) = I_S |\theta_S(\theta)| = I_S [\theta_S^0 + \mathcal{A}(\theta_0).(\theta - \theta_0)], \quad (7.27)$$

where we have expanded the relation $\theta_S(\theta)$ obtained from the lens equation at first order around $\theta_S^0 = \theta_0 - \alpha(\theta_0)$. This allows us to compare the differential effect on neighbouring light rays. The image distortion is thus characterized by the distortion matrix

$$\mathcal{A}(\theta) \equiv \frac{\partial \theta_S}{\partial \theta}, \quad (7.28)$$

the expression of which is obtained from the lens equation (7.20)

$$\mathcal{A}_{ab} = \delta_{ab} - \frac{\partial^2 \psi(\theta)}{\partial \theta^a \partial \theta^b}. \quad (7.29)$$

This symmetric (2×2) matrix involves 3 parameters, which can be decomposed in terms of the convergence, κ , and the shear, $\gamma = (\gamma_1, \gamma_2)$, as

$$\mathcal{A} = \begin{pmatrix} 1 - \kappa - \gamma_1 & \gamma_2 \\ \gamma_2 & 1 - \kappa + \gamma_1 \end{pmatrix}. \quad (7.30)$$

The convergence induces an isotropic focalization of the light beam and is related to the projected potential and to the surface density by

$$2\kappa(\theta) = \Delta_2\psi(\theta) = 2\frac{\Sigma}{\Sigma_{\text{crit}}}, \quad (7.31)$$

while the cosmic shear describes an anisotropic deformation and is related to the lens astigmatism. Its two components can be rewritten as

$$\gamma_1(\theta) = \frac{\psi_{,11} - \psi_{,22}}{2} \equiv \gamma(\theta) \cos[2\phi(\theta)], \quad \gamma_2(\theta) = \psi_{,12} \equiv \gamma(\theta) \sin[2\phi(\theta)]. \quad (7.32)$$

As represented in Fig. 7.3, an initially circular source of unit radius will have an elliptical image with minor and major axes $(1 - \kappa + \gamma)^{-1}$ and $(1 - \kappa - \gamma)^{-1}$, respectively, so that the image surface, S , is related to that of the source, S_0 , by

$$S = \mu S_0, \quad \mu = \frac{1}{\det \mathcal{A}} = \frac{1}{[(1 - \kappa)^2 - \gamma^2]}. \quad (7.33)$$

Since the differential deflection induces a variation in the shape and the surface of the source, it also modifies the light flux that is proportional to its surface. \mathcal{A} is thus the inverse of the amplification matrix, \mathcal{M} , which generalizes the expression (7.9) and we have

$$\mu = \frac{1}{\det \mathcal{A}} = \det \mathcal{M}. \quad (7.34)$$

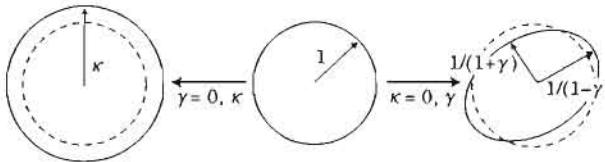


Fig. 7.3 Distortion of a circular source under the effect of the convergence κ and the cosmic shear γ .

Notice that \mathcal{A} is different for different multiple images. The comparison of the relative fluxes can in principle be used to constrain the lens matter distribution. For instance, in the case of a point-like lens, when $\theta_S \ll \theta_E$, the second image is highly attenuated and is usually undetectable. The quantity μ can be positive as well as negative. Its sign is called parity.

In conclusion, the distortion field of a gravitational lens in the thin-lens regime is characterized by the matrix

$$\mathcal{A} = [1 - \kappa(\theta)] \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \gamma(\theta) \begin{pmatrix} \cos[2\phi(\theta)] & \sin[2\phi(\theta)] \\ \sin[2\phi(\theta)] & -\cos[2\phi(\theta)] \end{pmatrix}, \quad (7.35)$$

where the projected potential is related to the surface density $\Sigma/\Sigma_{\text{crit}} = \kappa$ by

$$\psi(\theta) = \frac{1}{\pi} \int d^2\theta' \kappa(\theta') \ln |\theta - \theta'|. \quad (7.36)$$

Different regimes

The amplitude of the lensing depends on $\Sigma/\Sigma_{\text{crit}}$. Two main regimes can be distinguished.

- If $\Sigma/\Sigma_{\text{crit}} > 1$, the lens is said to be *supercritical* or *strong*. The convergence and the cosmic shear are important. Giant arcs as well as multiple images can appear.
- If $\Sigma/\Sigma_{\text{crit}} < 1$, we are in a regime of weak distortion; $\kappa \ll 1$ and $\gamma \ll 1$. There are no multiple images and the distortion of the background sources is so weak that it cannot be detected with the naked eye.
- Outside the critical zone, the images can also be subject to strong distortions without necessarily producing multiple images. This intermediate regime is called the *arclet regime*.

Figure 7.4 describes the transition between the two regimes in the case of a point-like lens. If the source is far from the Einstein radius, only its shape is affected and all the more so as it gets closer to the deflector. When it enters the Einstein radius, multiple images appear, and then giant arcs before finally entirely covering the Einstein radius when the source coincides with the optical axis. Figure 7.5 illustrates these different regimes in the case of the galaxy cluster Abell 2218.

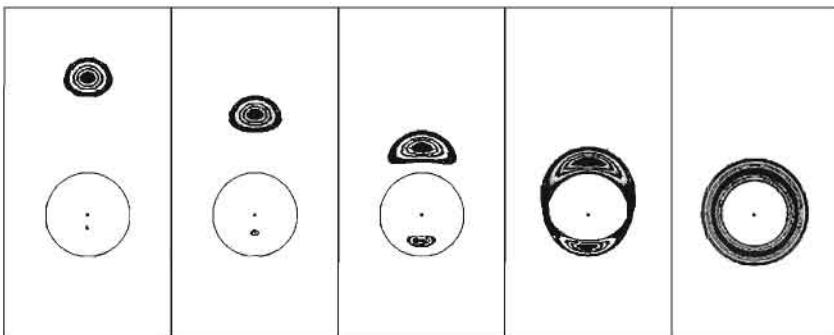


Fig. 7.4 Distortion of an extended source by a point-like lens. The circle represents the Einstein radius. The multiple images and the strong distortion effects only appear when the source is inside the Einstein radius. From Ref. [3].

The closed lines on the image plane for which $\det \mathcal{A} = 0$ are called *critical lines*. They characterize the positions of the strong magnification zones. The locus of the points of the source plane that have their image on the critical line is called a *caustic*. It indicates the position of the sources that will be subject to strong lensing and form arcs. One can show that when the source crosses a caustic, the number of images jumps by ± 2 , so that two images appear or blend. A close source inside a caustic produces two very bright images close to the critical line. A source close to a cusp produces three images (see the illustration in Fig. 7.7).

For a lens with spherical symmetry, the critical line is circular. For a point mass, it coincides with the Einstein radius and the associated caustic reduces to a point at



Galaxy Cluster Abell 2218

HST - WFPC2

NASA, A. Fruchter and the ERO Team (STScI, ST-ECF) · STScI-PRC00-08

Fig. 7.5 The galaxy cluster Abell 2218 provides especially spectacular lensing effects. The different regimes described in the text can be observed: giant arcs and multiple images close to the heart of the cluster, strong distortions without multiple images (arclets) and weak distortion effects for sources far away from the deflector. (Picture from the spatial Hubble telescope, © A. Fruchter, HST/WFPC2).

the intersection of the optical axis and the source plane (see Fig. 7.1).

One can also show [5] that if the function Σ is smooth, bounded (so that α does not diverge), and non-singular (so that α is continuous), then the number of images of (almost) all the sources is odd. Moreover, if Σ never becomes negative, then at least one of the images is amplified. In general, these relations and theorems are violated by observations, mainly due to the existence of many substructures in the lenses. Furthermore, the very damped images ($\kappa \ll 1$) cannot be detected in general.

Time delay

Using the relation (7.24), the lens equation (7.14) can be rewritten as

$$\nabla_2 \left[\frac{1}{2} (\theta - \theta_S)^2 - \psi \right] = 0,$$

where the term that appears in brackets is proportional to the time delay

$$\tau(\theta, \theta_S) = \frac{1+z_L}{c} \frac{D_{OL} D_{OS}}{D_{LS}} \left[\frac{1}{2} (\theta - \theta_S)^2 - \psi \right], \quad (7.37)$$

between the real trajectory and the unperturbed trajectory. The first term, quadratic in $(\theta - \theta_S)$, has a purely geometrical origin and reflects the length difference between

the real light path and the unperturbed one. The second term has a relativistic origin and is identical to the Shapiro effect (see Chapter 1), $c\delta t = \int 2(1+z)\Phi dz/c^2$ where the factor $(1+z)$ takes into account the cosmic expansion; it is then approximatively $(1+z_L)$ in the thin-lens approximation. All the equations we have obtained can be derived by looking for the extremum of the function $\tau(\theta, \theta_S)$. Figure 7.6 compares this approach to that based on the lens equation. This function has various interesting properties.

- The difference $\tau(\theta_1, \theta_S) - \tau(\theta_2, \theta_S)$ of the time delays between two stationary points gives the delay in the reception of the two images. Any variability in the source is first observed in the image corresponding to the smallest value of $\tau(\theta, \theta_S)$.
- There are three types of stationary points, maxima, minima and saddle points, the nature of which depends on the sign of the eigenvalues of the matrix $\mathcal{T} = \partial^2 \tau / \partial \theta_a \partial \theta_b$, which is simply the distortion matrix. It therefore describes the local curvature of isotime surfaces. We can thus define three types of images.
 - Type I: images corresponding to minima of \mathcal{T} . Both eigenvalues are positive and thus $\det \mathcal{A} > 0$ and $\text{Tr } \mathcal{A} > 0$.
 - Type II: images corresponding to saddle points of \mathcal{T} . The two eigenvalues have opposite sign, and thus $\det \mathcal{A} < 0$.
 - Type III: images corresponding to maxima of \mathcal{T} . Both eigenvalues are negative, and thus $\det \mathcal{A} > 0$ and $\text{Tr } \mathcal{A} < 0$.
- Images of type I and II have a positive parity and only type III images have a negative parity.

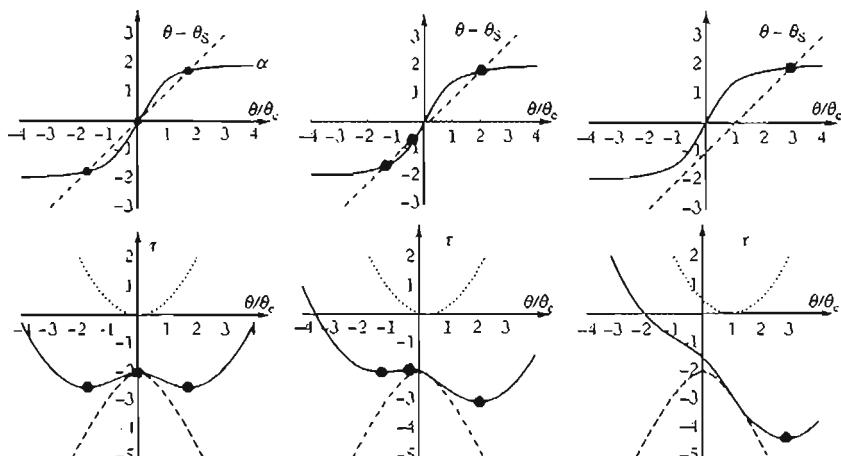


Fig. 7.6 Number of images for a smooth isothermal lens determined by the lens equation (top) or by the time delay (bottom) for a source localized at $\theta_S/\theta_c = 0, 0.3, 1$ from left to right. The dotted-line parabolae represent the geometrical contribution, while the other dotted curve is the gravitational contribution.

7.1.1.3 Lens models

The separation between the multiple images and the position of the critical lines enables us to reconstruct the surface density distribution of the lens. The aim is to adjust this distribution and the position of the source so that $\theta_S = \theta_i - \alpha(\theta_i)$ for all images θ_i .

We present here some lens models that illustrate the effects discussed in the previous sections.

7.1.1.4 Singular isothermal sphere

A simple model for the mass distribution of a galaxy relies on the hypothesis that stars behave as a perfect gas confined in a gravitational potential. Its equation of state will therefore be $P = \rho kT/m$, where m is the mass of each star, assumed to be all equal, and $\rho = Nm/V$. The temperature is related to the velocity dispersion by $m\sigma_v^2 = kT$. For an isothermal gas, T is uniform in the gas so that the hydrostatic equilibrium equation implies that $dP/dr = -G_N M(r)\rho/r^2$, with $dM/dr = 4\pi r^2 \rho$, assuming spherical symmetry. As a result, the density profile of such a galaxy is

$$\rho(r) = \frac{\sigma_v^2}{2\pi G_N} \frac{1}{r^2}. \quad (7.38)$$

This distribution is singular at the centre and the total mass contained in a sphere of radius r , $M(r)$, grows as r , so that the Keplerian velocity of a test particle in a circular orbit at distance r from the centre is

$$v^2(r) = \frac{G_N M(r)}{r} = 2\sigma_v^2, \quad (7.39)$$

which is constant. We can also conclude that the surface density is

$$\Sigma(b) = \frac{\sigma_v^2}{2G_N} \frac{1}{b}. \quad (7.40)$$

Using (7.26) and (7.24), we find that the deflection angle is independent of the impact parameter and is given by

$$\alpha = 4\pi \frac{\sigma_v^2}{c^2} = 1.4 \left(\frac{\sigma_v}{220 \text{ km} \cdot \text{s}^{-1}} \right)^2 \text{ arcsec.} \quad (7.41)$$

The Einstein radius is obtained by solving (7.14) for $\theta_S = 0$ and is $\theta_E = (D_{LS}/D_{OS})\alpha$.

The lens equation (7.14) takes the simple form $\theta_{\pm} = \theta_S \pm \theta_E$ and there are multiple images only if the source is inside the Einstein ring ($\theta_S < \theta_E$) and their angular separation is always $\Delta\theta = 2\theta_E$. Actually, there is a third image with vanishing flux at $\theta = 0$ that becomes visible if the profile is smoothed out at $r = 0$. The magnification (7.33) of the two images is $\mu_{\pm} = (1 \mp \theta_E/\theta_{\pm})^{-1}$. If $\theta_S > \theta_E$ then there is only one image localized at $\theta = \theta_S + \theta_E$. Figure 7.6 compares the graphic resolution of this system from the lens equations and that obtained by looking for the extremum of the time delay.

7.1.1.5 Other models of halos

The singular isothermal model is not realistic for various reasons. In particular, the lens must have a core where the density does not diverge. Table 7.2 summarizes the optical properties of some frequently used models with spherical symmetry.

Table 7.2 Properties of some models of spherically symmetric halos.

| Lens model | $\psi(\theta)$ | $\alpha(\theta)$ |
|----------------------------|---|--|
| Point-like mass | $(4G_N M / Dc^2) \ln \theta$ | $(4G_N M / c^2 D_{OL}) \theta^{-1}$ |
| Singular isothermal sphere | $(D_{LS} / D_{OS}) 4\pi (\sigma_v^2 / c^2) \theta$ | $4\pi \sigma_v^2 / c^2$ |
| Smoothed isothermal sphere | $(D_{LS} / D_{OS}) 4\pi (\sigma_v^2 / c^2) (\theta_c + \theta^2)^{1/2}$ | $4\pi (\sigma_v^2 / c^2) \theta / (\theta_c^2 + \theta^2)^{1/2}$ |
| Mass sheet | $\kappa \theta^2 / 2$ | $(D_{OS} / D_{LS}) \kappa \theta$ |

These models are still too idealized to describe real galaxies. In particular, one should go beyond spherically symmetric models. For instance, one can use elliptical models [6] for which the surface density takes the form

$$\Sigma(\theta) = \Sigma_0 [\theta_c^2 + (1 - \epsilon)\theta_1^2 + (1 + \epsilon)\theta_2^2]^{-1/2}. \quad (7.42)$$

In general, the projected potential is difficult to obtain. We can therefore choose to directly model this potential by assuming that it is elliptical

$$\psi(\theta) = \frac{D_{LS}}{D_{OS}} 4\pi \frac{\sigma_v^2}{c^2} [\theta_c^2 + (1 - \epsilon)\theta_1^2 + (1 + \epsilon)\theta_2^2]^{1/2}. \quad (7.43)$$

Figure 7.7 represents the critical and caustic lines for a model of an elliptic lens.

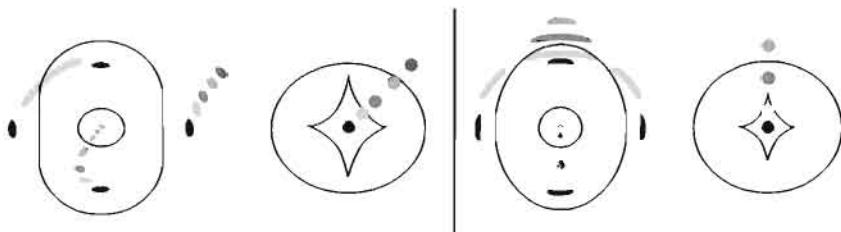


Fig. 7.7 Two examples of gravitational lenses. For each one of them, the figure on the left represents the image plane with the critical lines and the figure on the right the source plane with the caustic lines and the different source positions. From Ref. [4].

7.1.1.6 Effect of the environment

A galaxy is not isolated, in general it belongs to a cluster or group that induces an external gravitational distortion of the order of a few per cent. Another problem arises

due to the mass-sheet degeneracy. Let us add a mass sheet of constant density while performing the transformation

$$\kappa \rightarrow (1-s) + s\kappa, \quad \theta_S \rightarrow s\theta_S. \quad (7.44)$$

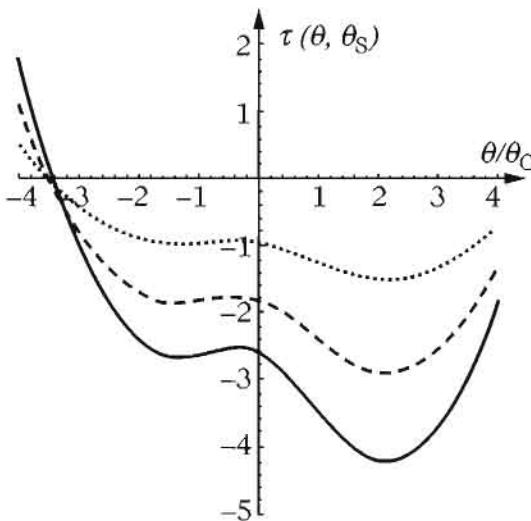


Fig. 7.8 Change of the function $\tau(\theta, \theta_S)$ for a smoothed isothermal profile when adding a mass sheet characterized by $s = 1$ (dashes) $s = 0.5$ (dots) and $s = 1.5$ (solid line). The angular positions of the maxima and minima remain unchanged.

The deflection angle then transforms as $\alpha \rightarrow s\alpha + (D_{\text{OS}}/D_{\text{LS}})(1-s)\theta$, so that the lens equation (7.20) remains unchanged since it is only multiplied by the global factor s . In an analogous way, neglecting the term θ_S^2 from the time delay (7.37) (in any case it is not observable since only differences $\delta\tau_{12} = \tau(\theta_1, \theta_S) - \tau(\theta_2, \theta_S)$ between various images can be measured) and noticing that the projected potential transforms as $\psi \rightarrow s\psi + (1-s)\theta^2/2$, then $\delta\tau_{12} \rightarrow s\delta\tau_{12}$. All the time delays are thus dilated in the same way, conserving the structure of the image. Since the source plane is dilated by s , all magnifications are damped by $1/s$, thus leaving the relative magnifications unchanged (see Fig. 7.8).

7.1.2 Lensing by galaxies and galaxy clusters

This section describes some applications of gravitational lensing in the thin-lens regime. For more details, we refer the reader to Refs. [7,8] for microlensing, to Ref. [9] for giant arcs and the arclet regime in clusters of galaxies, to Refs. [3,10] for lensing on quasars, and to Refs. [4,11,12] for comprehensive reviews of all observations.

7.1.2.1 Microlenses in our galaxy

If a massive body passes in front of a background source, it can induce a temporary variation of its apparent luminosity even if the mass of the lens is too small to induce

any multiple images or any important deflections. This phenomena is then called *microlensing*.

The estimate (7.4) shows that a lens of a few solar masses has an Einstein radius of a few milliarcseconds, which is well below the spatial resolution of our instruments (even if multiple images were produced, we would not be able to distinguish each one of them). However, if the source passes near the Einstein radius ($\theta_S \simeq \theta_E$), the total magnification (7.11) would be $\mu \simeq 1.34$. So, if the impact parameter of the source is smaller than θ_E , the peak of the light curve is $\mu_{\max} > 1.34$. This corresponds to a magnitude variation greater than 0.32 magnitudes, which is in principle detectable.

Paczniński [7] had the idea to use these microlensing phenomena to constrain the density of MACHOs (massive astrophysical compact halo object) in our Galaxy. The characteristic time for the evolution of the luminosity of a source depends on the Einstein radius and on the lens velocity, v ,

$$\tau_* = \frac{D_{OL}\theta_E}{v}. \quad (7.45)$$

Using the estimate (7.4) for the Einstein radius, we get

$$\tau_* = 0.214 \left(\frac{M}{M_\odot} \right)^{1/2} \left(\frac{D_{OL}}{10 \text{ kpc}} \right)^{1/2} \left(\frac{D_{LS}}{D_{OS}} \right)^{1/2} \left(\frac{v}{200 \text{ km} \cdot \text{s}^{-1}} \right)^{-1} \text{ years}. \quad (7.46)$$

If the lens is located in our Galaxy and the source in the large Magellanic cloud, then D_{LS}/D_{OS} is of order unity. Then, a sampling of light curves with time intervals ranging between an hour and a day makes it possible to detect masses between $10^{-6} M_\odot$ and $10^2 M_\odot$. We should, however, stress that the measurement of τ_* does not directly give access to the lens mass.

7.1.2.2 Light curves

The light curve of the source as a function of time can be obtained using (7.11) for the amplification by a point-like source,

$$A(t) = \frac{2 + x^2}{x\sqrt{4 + x^2}}, \quad (7.47)$$

where $x(t) \equiv \sqrt{u_{\min}^2 + a^2 t^2}$, u_{\min} being the impact parameter in units of the Einstein radius and a is related to the characteristic time (7.46). Such light curves are depicted in Fig. 7.9 for various impact parameters.

Both parameters, a and u_{\min} , can thus be determined by fitting an observed light curve. The width of the curve gives access to τ_* and there is a relation between the amplitude of the peak and u_{\min} . This is, however, not sufficient to determine all of the characteristics of the system since there are three unknowns (M , D_{OL} and v).

7.1.2.3 Cross-section

To evaluate the probability of a microlensing, it is convenient to define a cross-section as the surface around the lens through which the light emitted by the source must

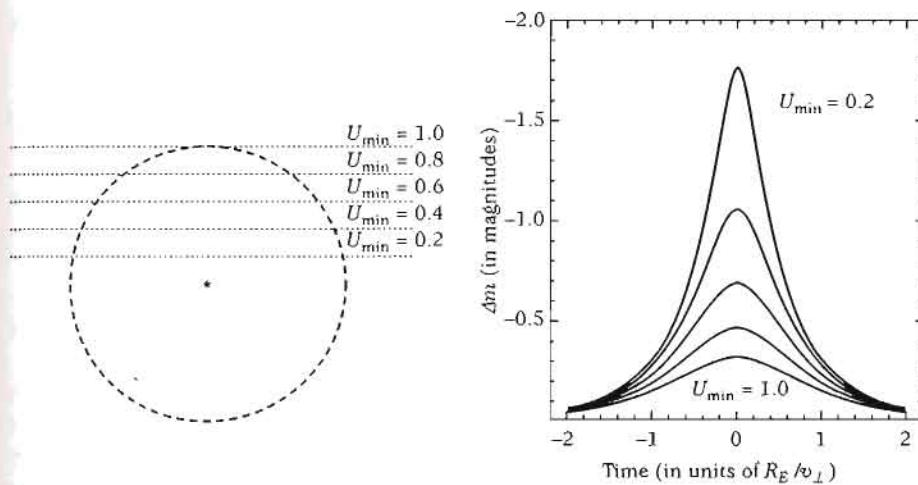


Fig. 7.9 Light curves during a microlensing event for different values of the impact parameter in units of the Einstein radius.

pass in order to be subject to a magnification larger than A . For a point-like source, this surface is $\pi D_{\text{OL}}^2 \theta_S^2 (\mu = A)$ and thus

$$\sigma(>A) = 2 \frac{M}{\Sigma_{\text{crit}}} \frac{1}{A^2 - 1 + A\sqrt{A^2 - 1}}, \quad (7.48)$$

using (7.11). The number of lenses with a magnification greater than A is thus given by

$$N(>A) = \frac{1}{\Delta\Omega} \int_{< D_{\text{OL}}} \int_{\Delta\Omega} dV n(D_{\text{OL}}) \sigma(>A). \quad (7.49)$$

This quantity can be expressed in terms of the optical depth τ , defined as the integral of the area contained inside the Einstein ring of all the lenses in a solid angle $\Delta\Omega$,

$$\tau = \frac{1}{\Delta\Omega} \int_{< D_{\text{OL}}} \int_{\Delta\Omega} dV n(D_{\text{OL}}) \pi \theta_E^2. \quad (7.50)$$

Since $dV = d\Omega^2 D_{\text{OL}}^2 dD_{\text{OL}}$ and expressing θ_E , this expression reduces to

$$\tau = \frac{4\pi G_N}{c^2} \int_0^{D_{\text{OL}}} \rho_{\text{macho}}(D_{\text{OL}}) \frac{D_{\text{OL}} D_{\text{LS}}}{D_{\text{OS}}} dD_{\text{OL}}, \quad (7.51)$$

where ρ_{macho} is the density of MACHOs. By introducing $x = D_{\text{OL}}/D_{\text{OS}}$, we finally obtain

$$\tau = \frac{4\pi G_N D_{\text{OS}}^2}{c^2} \int_0^1 \rho_{\text{macho}}(x) x(1-x) dx. \quad (7.52)$$

We have assumed that since space is Euclidean at these scales, we can make the approximation $D_{\text{LS}} = D_{\text{OS}} - D_{\text{OL}}$. For a constant density, we will thus obtain $\tau =$

$(2\pi/3)G_N \rho_{\text{macho}} D_{\text{OS}}^2/c^2$. The theoretical estimates predict that $\tau \sim 10^{-6}$. The number of events that produce a magnification greater than A is thus given by

$$N(> A) = \frac{2\tau}{A^2 - 1 + A\sqrt{A^2 - 1}}, \quad (7.53)$$

for $A = 1.34$, $N \sim \tau$ and several millions of stars should be monitored before one can hope to observe any event.

7.1.2.4 Searching for MACHOs

Several experiments have carried out this idea in order to constrain the density of MACHOs in the halo of our Galaxy.

The main problem resides in the amount of data needed (several millions of stars are observed) and in the detections that are not associated with any lensing effect. In particular, an automated monitoring will detect many variable stars and supernovæ. Two criteria can be used to distinguish between the magnification induced by lensing from such intrinsic variations: the light curve must be symmetric and the magnification must be achromatic.

Figure 7.10 summarizes the results from the two experiments MACHO [13] and EROS [14] that have been looking for microlensing effects on the stars of the large Magellanic cloud.

7.1.2.5 Gravitational lensing by galaxies

Quasars

The first multiple quasar, QSO0957+561 ($z = 1.41$), was discovered in 1979 [16] and today we know of more than 10^4 of them. Microlensing was initially observed in the four images of the quasar QSO2237+0305 [17] in 1989. The deflector is a spiral galaxy at redshift $z \sim 0.04$. Uncorrelated variations of the luminosity from the four images allow us to establish that the microlensing is induced by stars from the galaxy.

In this situation, a complex caustic network appears and can be used to put constraints on the structure and the intrinsic size of the source. Moreover, one can determine the mass spectrum of these microlenses and thus show that its star distribution is analogous to that of our Galaxy. Let us note some of the results obtained by these observations.

- The central image is absent in almost all the observations. This implies that it is very damped and that almost all galaxies must have a very small core radius, $r_c < 200$ pc.
- One can show [18] that most galaxies must have a dark-matter halo with a velocity dispersion of the order of 220 ± 20 km·s $^{-1}$. If this were not the case, no multiple images with angular separation greater than 2" would be produced. The largest estimated size for a halo is that of the galactic lens QSO0957+561 that is greater than $15h^{-1}$ kpc.
- The optical depth depends on the redshift distribution of the lenses. If galaxies were formed recently, its main contribution would be from galaxies at low redshift.

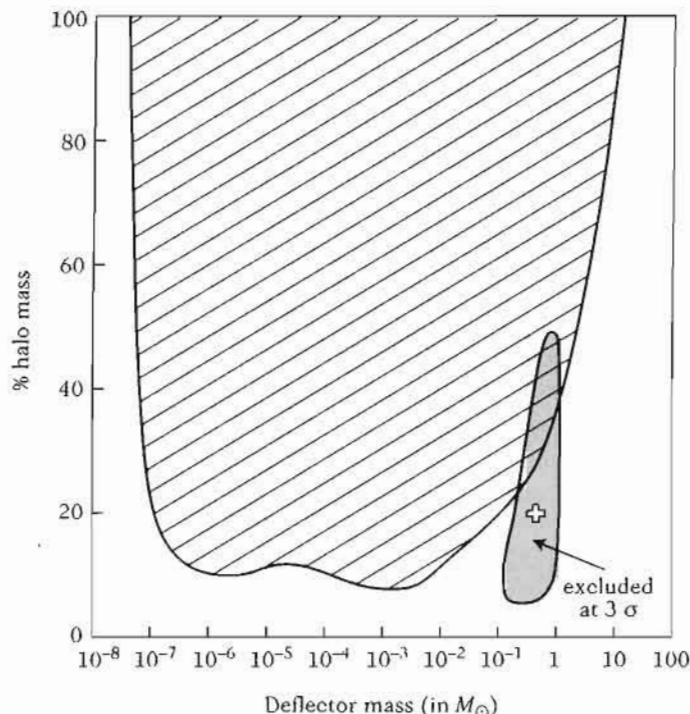


Fig. 7.10 Constraints stemming from the results of the MACHO (grey tint) and EROS (hatched) experiments on the contribution of MACHOs to the halo of our Galaxy (modelled by a mass of $4 \times 10^{11} M_\odot$ and a radius of 50 kpc). The allowed region is the one below the U curve. From Ref. [15].

Using the information on the lenses and sources redshift, one can conclude that most galaxies must have been formed before $z \sim 0.8$ assuming an Einstein-de Sitter Universe model [19].

- The standard cold dark-matter model predicts the formation of dark halos, which tend to induce more pairs of quasars with large separations than that observed. This could be a potential problem for this scenario but the conclusion depends on the properties of these halos (core radius...) and on the power spectrum of dark matter [20]. This illustrates the potential for this tool to test the distribution of dark matter.
- A cosmological constant has the effect of increasing dV/dz at large redshifts. This implies [21] that the relative number of lensed sources must rapidly increase with Λ , which can thus be estimated using the lens probability. The current upper bound obtained from this method is $\Omega_{\Lambda 0} < 0.65$ (2σ) for a flat Universe [22]. Another approach [23], using the comparison between the observed separations between multiple quasars and that expected given the sources redshifts, has led to $0.48 < \Omega_\Lambda < 0.93$.

- A last application, not yet exploited, concerns the differential magnification by stars from the bulge, of small Einstein radius, of the quasar accretion disk. This effect might allow us to probe the disk and to study its structure.

Hubble constant

The measurement of the time delay between different images of the same source gives access to the Hubble constant, as noted by Refsdal [24], since the time delay $\tau(\theta, \theta_S)$ is proportional to H_0 .

To understand this, let us consider the model of an isothermal lens. The lens equation gives access to $\theta - \theta_S$ (which is proportional to the deflection angle) for each image. We obtain that $\tau(\theta, \theta_S) \propto \dot{\alpha}^2(\theta)/2 - \psi(\theta)$. For an isothermal lens, the deflection angle is constant so that the first term does not contribute and $\delta\tau \propto \psi(\theta_1) - \psi(\theta_2) \propto \theta_1 - \theta_2 = 2\theta_E$. We finally deduce that

$$c\delta\tau = 8\pi \left(\frac{\sigma_v^2}{c^2} \right)^2 (1 + z_L) D_{OL} \theta_E. \quad (7.54)$$

In this equation, only D_{OL} is proportional to H_0^{-1} . Of course, the determination of H_0 depends on the lens model and is very sensitive to the existence of a mass-sheet (see Ref. [25] for a modellisation of these effects).

For instance, for the quasar QSO0957+561, one can show [26] that

$$h = (0.82 \pm 0.06)(1 - \kappa) \left(\frac{\Delta\tau}{1.14 \text{ year}} \right)^{-1}, \quad (7.55)$$

where κ is the unknown convergence of the halo containing the galaxy. To measure such time delays, one should therefore monitor these galaxies for several years. For QSO0957+561, the time delay is

$$\delta\tau = (417 \pm 3) \text{ days}. \quad (7.56)$$

Since κ is positive, we deduce an upper bound for the Hubble constant ($h < 0.88$). κ can be estimated from the weak lensing, giving $\kappa = 0.24 \pm 0.12$, which allows us to conclude [27] that $h = 0.49 - 0.74$. The measurements of the galaxy velocity dispersion can also be used to constrain κ and gives [28] $h = 0.66 \pm 0.07$.

This method requires that the geometry of the lens and that of its environment are well understood. It has, however, some advantages. For instance, it avoids the construction of a distance scale calibrated step by step and to work with sources at large redshifts ($z \sim 0.5$). If the deflector potential is perfectly known, the result should not depend on the physics of the particular object (cepheids, supernovæ...), and the only underlying hypothesis is the validity of general relativity in the weak-field limit.

Constraints on the density of compact objects

It has been suggested that a significant part of dark matter could be in the form of compact objects. The density of such a component of dark matter can be constrained from microlensing effects.

For this, one should estimate the optical depth (7.50) in the cosmological context. It can no longer be assumed that $D_{LS} = D_{OS} - D_{OL}$ and one should use the expressions obtained in Chapter 3. Assuming a constant density ρ_M , we obtain

$$\tau(z_S) = \frac{3\Omega_{M0}}{2} \int_0^{z_L} \frac{D_{LS}(z, z_S)}{D_{OS}(z_S)} \frac{D_{OL}(z)}{D_{H_0}} \frac{dz}{E(z)}, \quad (7.57)$$

where we have used the fact that $dV(z) = D_{OL}^2 d\chi d\Omega^2 / (1+z)^2$ and $d\chi = dz/H(z)$. For an Einstein-de Sitter Universe, we obtain

$$\tau(z_S) = 3 \frac{\Omega_{M0}}{z_S} [(z_S + 2 + 2\sqrt{1+z_S}) \ln(1+z_S) - 4z_S]. \quad (7.58)$$

This rate is proportional to Ω_{M0} so that the number of events in a given catalogue is a direct estimate of the value of this parameter. Different analyses can be used to exclude $\Omega_{M0} = 1$ and give the constraint $\Omega_{M0} < 0.05$ for masses $M = 10^6 - 10^{12} M_\odot$.

7.1.2.6 Lensing by galaxy clusters

The first giant arcs were discovered in 1986 in the clusters Abell 370 [29], Abell 2218 and Cl 2214 [30]. The numerous properties of the cluster distortion field allow us to reconstruct its gravitational potential and thus to study its structure.

Reconstruction of the dark-matter distribution

Together with the analyses of the velocity dispersion and the X-ray emission, gravitational lenses provide an independent method to determine the total mass of clusters.

Strong lensing mainly appear in the central zones of the cluster ($\lesssim 50 h^{-1}$ kpc) and can be used to measure the internal mass. A first estimate of this mass is given by $M(< \theta_E)$ since, for a spherical lens, the position of the critical lines gives a measure of the Einstein radius. This estimate has a precision of the order of 50% and strongly depends on the degree of asymmetry of the cluster. When several arcs and multiple images are present (see Fig. 7.5), the modelling of the cluster is refined. One can then proceed by iteration: once a first model is obtained, the position of other multiple images can be predicted and their presence can be checked. In this case, one can deduce the expected morphology deduced from the model and iterate it. These models become precise enough to determine the redshifts of the arcs and arclets [31].

The hot gas of the cluster emits Bremsstrahlung radiation with intensity depending on the ions (mainly protons) and electron number density, and on the gas temperature as well as on its chemical composition. From the measurements of the X-ray emission and the temperature profile, one can reconstruct the cluster mass profile, $M(r)$. The discrepancy between X-ray and lensing measurements is mainly attributed to the influence of the substructures on the X-ray emission and to the triaxiality of the three-dimensional potential.

These methods have allowed us to determine generic properties of the dark-matter distribution in clusters, independently of any given model. Clusters seem to be dominated by dark matter and their typical mass-luminosity ratio would be $M/L > 100 M_\odot/L_\odot$. The distribution of dark matter seems to follow that of the luminous

one, in particular in the central regions. The curvature radius of many giant arcs is comparable to their distance to the centre of the cluster, which implies that the cluster's core radius should be of the same order of magnitude. The internal zones of the clusters have many substructures.

Magnification bias

The magnification bias expresses two simultaneous and opposite effects to the gravitational magnification: it increases the flux received from the galaxies that are subject to lensing and so allows for the detection of a larger number of objects, but it also dilates the considered solid angles of the same quantity and thus decreases the density of galaxies.

If $n_0(f, z)$ is the number of galaxies in the absence of lenses per unit solid angle with flux greater than f at redshifts between z and $z + dz$, then the density of galaxies in a direction θ becomes, because of magnification,

$$n(f, z) = \frac{1}{\mu(\theta, z)} n_0 \left[\frac{f}{\mu(\theta, z)}, z \right]. \quad (7.59)$$

The magnification can thus either locally increase or decrease the galaxy count. If we only have access to the number count integrated over z , we can make the simplifying hypothesis that $\mu(\theta, z) \sim \mu(\theta)$. Assuming that the galaxy number count as a function of the magnitude has a slope $\alpha \equiv d \log n_0 / dm$ then

$$n(m, \theta) = n_0(m) |\mu(\theta)|^{2.5\alpha - 1}. \quad (7.60)$$

An effect thus only appears if $\alpha \neq 0.4$. The variation of the object density as a function of the distance to the centre of the cluster therefore gives a depletion curve that allows us to measure the magnification and, for instance, to localize a critical line even if no giant arc is associated with it [32].

Cosmological parameters

Once a robust model of the cluster is constructed, one can use the information redundancy to measure the cosmological parameters using geometrical methods. The lens equations can always be decomposed as

$$\theta_i(z_{S,i}) = \theta_{S,i} - E(z_{S,i}) f(\theta_{i,j}, \dots) D_{OL}.$$

The function $E = D_{LS}/D_{OS}$ depends on the source redshift and f is a function that depends on the modelling of the cluster. If one has access to different sets of multiple images centred on the same gravitational lens, then one can estimate

$$\frac{E(z_{S,1})}{E(z_{S,2})} = \frac{|\theta_{1,1} - \theta_{1,2}|}{|\theta_{2,1} - \theta_{2,2}|} \frac{|f(\theta_{1,2}, \dots) - f(\theta_{2,2}, \dots)|}{|f(\theta_{1,1}, \dots) - f(\theta_{2,1}, \dots)|}. \quad (7.61)$$

We thus obtain a measurement of D_{LS}/D_{OS} at different redshifts. This method, which is simple in principle, is, however, difficult to apply in practice as the function f is not well known.

Detection of distant galaxies

A source close to a caustic can be enormously amplified. Giant arcs can thus allow for the observation of galaxies with low surface brightness at high redshift. The lens here plays the role of a gravitational telescope making it possible to observe objects that would otherwise be undetectable.

The information on the redshift at which the first structures of the Universe form, which can be obtained in this way, allows constraints to be placed on models of large-scale structure formation and on the cosmological parameters.

Currently several (5–10) galaxies with redshifts larger than 7 have been detected, amongst which one candidate has a redshift of 10. This galaxy [33], which is close to a critical line, has an estimated magnification of 25–100. In a flat Universe, with $\Omega_{\Lambda 0} = 0.7$ it would thus have been formed around 460 million years after the Big Bang.

7.1.3 Gravitational distortion by the large-scale structure

Up to now, we have only considered the effects of gravitational lensing in the thin-lens regime. Light beams emitted by galaxies are, nevertheless, continuously deformed and deflected by the gravitational field of the large-scale structures of the Universe. These effects are small and do not give rise to multiple images. However, the distortion of the shape of distant galaxies and the statistics of these distortions reflect some properties of large-scale structure.

7.1.3.1 Sachs equation

We start by studying the deformation of a geodesic bundle during its propagation through an inhomogeneous space-time [34]. As seen in Chapter 1, light follows null geodesics. Two geodesics of the same bundle will be subject to a relative deviation described by the geodesic deviation equation (1.118).

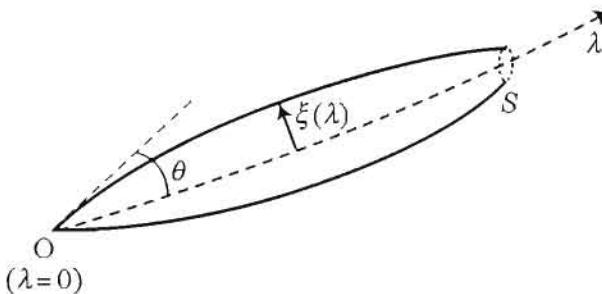


Fig. 7.11 Deviation of a geodesic bundle converging at the observer.

The different geodesics of a geodesic bundle can be parameterized as follows

$$x^\mu(\lambda) = \bar{x}^\mu(\lambda) + \xi^\mu(\lambda), \quad (7.62)$$

where \bar{x}^μ is the worldline of an arbitrary geodesic of the bundle chosen as reference and ξ^μ is a vector characterizing the displacement between two neighbouring geodesics (see Fig. 7.11); λ is an affine parameter along the reference geodesic. We assume that it is zero at O (position of the observer) and increases towards the past.

In O , we can choose an orthonormal basis, $\{k^\mu, u^\mu, n_1^\mu, n_2^\mu\}$ composed of the tangent vector to the reference null geodesic,

$$k^\mu \equiv \frac{d\bar{x}^\mu}{d\lambda},$$

of the tangent vector to the observer worldline, u^μ , and of two space-like vectors in the plane orthogonal to the line of sight. They thus satisfy

$$k^\mu k_\mu = 0, \quad u^\mu u_\mu = -1, \quad n_a^\mu n_b^\mu = \delta_a^b, \quad n_a^\mu k_\mu = n_a^\mu u_\mu = 0, \quad (7.63)$$

with $a, b = 1, 2$. From this basis defined in O , one can construct a basis at every point on the reference geodesic by parallel transport (that is according to $k^\mu \nabla_\mu n_\nu^a = 0$ and $k^\mu \nabla_\mu u_\nu = 0$). Any vector ξ^μ can then be decomposed as¹

$$\xi^\mu = \xi_0 k^\mu + \sum_a \xi_a n_a^\mu, \quad (7.64)$$

and one can always set $\xi_0 = 0$ since two different values of ξ_0 still parameterize the same geodesics. The propagation equation for ξ^μ follows from the geodesic deviation equation (1.118)

$$\frac{D^2}{d\lambda^2} \xi^\mu = R^\mu_{\nu\alpha\beta} k^\nu k^\alpha \xi^\beta, \quad (7.65)$$

where $R^\mu_{\nu\alpha\beta}$ is the Riemann tensor and $D/d\lambda \equiv k^\mu \nabla_\mu$. Using the decomposition (7.64), this equation becomes

$$\frac{d^2}{d\lambda^2} \xi = \mathcal{R} \cdot \xi, \quad \mathcal{R}_a^b = R^\mu_{\nu\alpha\beta} k^\nu k^\alpha n_{a\mu} n^{b\beta}, \quad (7.66)$$

where we have used the notation $\xi = \xi_a$ and $\mathcal{R} \cdot \xi = \mathcal{R}_a^b \xi_b$. Here a and b are just labels not tensor indices. The optical matrix, \mathcal{R} , can be decomposed in terms of the Ricci and the Weyl tensors as

$$\mathcal{R}_a^b = -\frac{1}{2} R_{\mu\nu} k^\mu k^\nu \delta_a^b + C^\mu_{\nu\alpha\beta} k^\nu k^\alpha n_{a\mu} n^{b\beta}. \quad (7.67)$$

The linearity of the geodesic equation implies that ξ is related to the initial value of its derivative by a linear transformation

$$\xi(\lambda) = \mathcal{D} \left. \frac{d\xi}{d\lambda} \right|_0. \quad (7.68)$$

¹A priori we should have added a term $A u^\mu$ in this decomposition. But, it can be shown that $k^\mu \xi_\mu = -A E$, with $E = -(k^\mu u_\mu)$, is constant along the geodesic. It follows that this term can be set to zero with no loss of generality.

The bundle converging in O , $\xi(0) = 0$ and (7.66) gives an evolution equation for the matrix \mathcal{D}

$$\frac{d^2}{d\lambda^2} \mathcal{D} = \mathcal{R} \cdot \mathcal{D}. \quad (7.69)$$

The initial conditions in O are then given by

$$\mathcal{D}(0) = \mathbf{I}, \quad \frac{d}{d\lambda} \mathcal{D}(0) = \mathbf{I}, \quad (7.70)$$

\mathbf{I} being the 2×2 unit matrix. The direction of observation, θ , and of the unlensed source, θ_S , are related to the displacement field by

$$\theta = \frac{d\xi}{d\lambda}(0), \quad \theta_S = \frac{\xi(\lambda_S)}{D_A(\lambda_S)}, \quad (7.71)$$

where λ_S is the value of the affine parameter at the source and where the distance D_A is by construction the angular distance. We then find from (7.68) that

$$\theta_S = \mathcal{A} \cdot \theta \quad \text{with} \quad \mathcal{A} = \frac{\mathcal{D}(\lambda_S)}{D_A(\lambda_S)}. \quad (7.72)$$

We recognize the deformation matrix (7.28) that can be decomposed, as previously, into a convergence κ and a cosmic shear, γ [see (7.30)].

7.1.3.2 Application to cosmology

The equations we have just derived did not depend on any hypothesis concerning space-time symmetry. We now apply them to the relevant cases for cosmology.

Homogeneous and isotropic spaces

For a Friedmann–Lemaître space-time, the metric takes the form (3.3) and is conformal to the static metric $\bar{g}_{\mu\nu} dx^\mu dx^\nu = -d\eta^2 + \gamma_{ij} dx^i dx^j$ with the notations of Chapter 3.

To solve (7.69) we use the fact that two conformal spaces have the same null geodesics (see Section 6.1) and will thus have the same optical matrix. Equation (7.69) then takes the simple form

$$\frac{d^2}{d\lambda^2} \mathcal{D} = -K \mathcal{D}, \quad (7.73)$$

where K is the curvature of the three-dimensional space with metric γ_{ij} [see (3.4)]. The solution of this equation is

$$\mathcal{D}^{(0)}(\lambda) = f_K(\lambda) \mathbf{I}, \quad (7.74)$$

f_K being defined by (3.5) so that $\mathcal{A}^{(0)} = \mathbf{I}$, where D_A has been expressed using (3.74).

Contributions from perturbations

To take into account the effects of inhomogeneities, we work in the framework of the cosmological perturbation theory presented in Chapter 5. The space-time metric

is assumed to be of the form $ds^2 = a^2(\eta)(\bar{g}_{\mu\nu} + h_{\mu\nu})dx^\mu dx^\nu$ and we expand the matrix \mathcal{D} in perturbations as $\mathcal{D} = \mathcal{D}^{(0)} + \mathcal{D}^{(1)} + \dots$, the first term being given by (7.74).

To first order, (7.69) reduces to

$$\frac{d^2}{d\lambda^2} \mathcal{D}^{(1)} + K \mathcal{D}^{(1)} = f_K(\lambda) \mathcal{R}^{(1)}(\lambda), \quad (7.75)$$

which has the integral solution

$$\mathcal{D}^{(1)}(\lambda) = \int_0^\lambda f_K(\lambda') f_K(\lambda - \lambda') \mathcal{R}^{(1)}(\lambda') d\lambda', \quad (7.76)$$

implying that the amplification matrix is given by

$$\mathcal{A}^{(1)}(\lambda) = \int_0^\lambda \frac{f_K(\lambda') f_K(\lambda - \lambda')}{f_K(\lambda)} \mathcal{R}^{(1)}(\lambda') d\lambda'. \quad (7.77)$$

Finally, one should compute $\mathcal{R}^{(1)}$ and go back to the initial space-time, the expanding one. The definition (7.66) implies that²

$$2\mathcal{R}^{(1)} = h_{\mu\nu,\alpha\beta} k^\mu k^\nu n^\alpha n^\beta. \quad (7.78)$$

Using now the decomposition (5.52) in Newtonian gauge for the scalar modes only, we obtain

$$\mathcal{R}^{(1)} = -\partial_{ab}(\Phi + \Psi) = -2\partial_{ab}\Phi, \quad (7.79)$$

where the second equality is only valid if the anisotropic pressure is negligible, which is a priori the case in the matter era [see (5.118)]. The amplification matrix in a cosmological space-time can thus be expressed in terms of the gravitational potential as³

$$\mathcal{A}_{ab} = \delta_{ab} - \partial_{ab}\psi(\theta, \chi), \quad \psi(\theta, \chi) \equiv \frac{2}{c^2} \int_0^\chi \frac{f_K(x') f_K(\chi - x')}{f_K(\chi)} \Phi[f_K(x')\theta, x'] dx'. \quad (7.81)$$

The distortion is thus obtained by integrating along the unperturbed geodesic and takes a form analogous to that obtained in the case of the thin-lens approximation, but introducing a projected potential that takes into account the entire matter distribution. Note that this is thus a first-order expression so that all the contributions coming from the lens-lens coupling are neglected. For most practical applications this is, however, an

²Here, we will make a simplifying hypothesis. Since we will be interested in modes smaller than the Hubble scale, we assume that the spatial curvature does not influence the perturbations and neglect it in the computation of $\mathcal{R}^{(1)}$, while we keep it in the geometrical factors.

³Note that this equation can be rewritten into the equivalent form

$$\mathcal{A}_{ab} = \delta_{ab} - \partial_{\theta_a \theta_b} \tilde{\psi}(\theta, \chi), \quad \tilde{\psi}(\theta, \chi) \equiv \frac{2}{c^2} \int_0^\chi \frac{f_K(x - x')}{f_K(\chi)f_K(x')} \Phi[f_K(x')\theta, x'] dx', \quad (7.80)$$

where the derivatives are now taken with respect to θ_a .

excellent approximation. The thin-lens regime can be recovered by assuming that the gravitational potential is localized on a lens plane at χ_L , $\Phi = \Phi[f_K(\chi)\theta, \chi]\delta(\chi - \chi_L)$. The convergence and the cosmic shear are obtained as

$$\begin{aligned}\kappa(\theta, \chi) &= \frac{1}{2}(\psi_{,11} + \psi_{22}) = \frac{1}{2}(\bar{\psi}_{,\theta_1\theta_1} + \bar{\psi}_{\theta_2\theta_2}), \\ \gamma_1(\theta, \chi) &= \frac{1}{2}(\psi_{,11} - \psi_{22}) = \frac{1}{2}(\bar{\psi}_{,\theta_1\theta_1} - \bar{\psi}_{\theta_2\theta_2}), \\ \gamma_2(\theta, \chi) &= \psi_{,12} = \bar{\psi}_{,\theta_1\theta_2}.\end{aligned}$$

7.1.4 Cosmic convergence

7.1.4.1 Effective convergence

The convergence, obtained from the trace of the amplification matrix, characterizes the density of effective matter integrated along the line of sight (7.25). We deduce from (7.81) and the Poisson equation (5.117) that it takes the form⁴

$$\kappa(\theta, \chi) = \frac{3}{2} \frac{H_0^2}{c^2} \Omega_{m0} \int d\chi' \frac{f_K(\chi') f_K(\chi - \chi')}{f_K(\chi)} \frac{\delta[f_K(\chi')\theta, \chi']}{a(\chi')} \quad (7.82)$$

If the sources have a redshift distribution $p_z(z)dz = p_\chi(\chi)d\chi$, then the effective convergence is obtained by weighting the convergence (7.82) with the source distribution as

$$\kappa(\theta) = \int p_\chi(\chi) \kappa(\theta, \chi) d\chi.$$

It then takes the final form⁵

$$\kappa(\theta) = \frac{3}{2} \frac{H_0^2}{c^2} \Omega_{m0} \int_0^{\chi_H} g(\chi) f_K(\chi) \frac{\delta[f_K(\chi)\theta, \chi]}{a(\chi)} d\chi, \quad (7.83)$$

with

$$g(\chi) = \int_\chi^{\chi_H} p_\chi(\chi') \frac{f_K(\chi' - \chi)}{f_K(\chi')} d\chi', \quad (7.84)$$

where χ_H is the value of χ associated with the size of the observable Universe. This expression shows that the convergence depends on the properties of the density contrast, on the cosmological evolution, but also on the total quantity of matter, Ω_{m0} .

7.1.4.2 Limber equation for the convergence power spectrum

We are mainly interested in the statistical properties of the convergence that can be related, thanks to (7.83), to that of the density contrast. As discussed in Chapter 5, the

⁴Here, we have split the 3-dimensional Laplacian to get $\Delta_\perp \Phi = \frac{3}{2} \Omega_{m0} H_0^2 \delta/a - \partial_{33} \Phi$. The term involving $\partial_{33} \Phi$ gives a vanishing contribution when the integral over χ' is computed.

⁵After we have used that $\int_0^{\chi_H} d\chi \int_0^\chi d\chi' = \int_0^{\chi_H} d\chi' \int_{\chi'}^{\chi_H} d\chi$, and then exchanged χ and χ' in our notations.

density field is a Gaussian stochastic variable whose statistical properties are described by its power spectrum [see (5.36)]. Just as the density contrast was expanded into Fourier modes as in (5.33), the convergence can be expanded into bidimensional Fourier modes as⁶

$$\kappa(\theta) = \int \frac{d^2\ell}{2\pi} \hat{\kappa}(\ell) e^{i\ell \cdot \theta}. \quad (7.85)$$

The convergence power spectrum can then be defined by the relation

$$\langle \hat{\kappa}(\ell) \hat{\kappa}(\ell') \rangle = \delta^{(2)}(\ell + \ell') P_\kappa(\ell). \quad (7.86)$$

The relation (7.83) expresses the convergence as a projection of the density contrast with weight $g(x)$ as $\kappa(\theta) = \int g(x) \delta[f_K(x)\theta, x] dx$. We can then easily convince ourselves that this implies that κ is also a homogeneous and isotropic Gaussian field. Its angular correlation function $\xi_\kappa(\theta) \equiv \langle \kappa(\varphi) \kappa(\varphi + \theta) \rangle$ then takes the form

$$\xi_\kappa(\theta) = \int d\chi d\chi' g(x) g(x') \langle \delta[f_K(x)\varphi, x] \delta[f_K(x')(\varphi + \theta), x'] \rangle. \quad (7.87)$$

The correlation function of the density field introduced in this integral can be computed by expanding δ as in (5.33) and takes the form

$$\int \frac{dk d^3 k'}{(2\pi)^3} e^{-ik_\perp \cdot \varphi f_K(x)} e^{-ik_3 x} e^{-ik'_\perp \cdot (\varphi + \theta) f_K(x')} e^{-ik_3 x'} \langle \delta(k, x) \delta(k', x') \rangle. \quad (7.88)$$

If we only consider small angles ($\theta \ll 1$) then $k_\perp^2 \gg k_3^2$ and the power is mainly carried by k_\perp . We can thus make the approximation

$$\langle \delta(k, x) \delta(k', x') \rangle \simeq \langle \delta(k_\perp, x) \delta(k'_\perp, x') \rangle \delta(k_3 + k'_3).$$

The integral over k_3 gives $2\pi\delta(x - x')$ so that after integration over x' we obtain

$$\xi_\kappa(\theta) = \int d\chi g^2(x) \int \frac{d^2 k_\perp}{(2\pi)^2} P_\delta(k_\perp, x) e^{ik_\perp \cdot \theta f_K(x)} dx. \quad (7.89)$$

Within this approximation, the convergence power spectrum is therefore given by⁷

$$P_\kappa(\ell) = \frac{9H_0^4}{4c^4} \Omega_{m0}^2 \int_0^{x_H} \left[\frac{g(x)}{a(x)} \right]^2 P_\delta \left[\frac{\ell}{f_K(x)}, x \right] dx. \quad (7.91)$$

Such an expression allows us to predict the amplitude of the convergence as a function of the angular scale and of the cosmological model at hand.

⁶We are still in a flat-sky approximation. In full generality, $\kappa(\theta)$ is a function on the celestial sphere and should be expanded in spherical harmonics, as, e.g., the CMB temperature anisotropies. For more details on this approach and on its relation to the flat-sky limit, see, e.g., Ref. [35].

⁷Note that if we had not used the Poisson equation, the same derivation would have given

$$P_\kappa(\ell) = \int_0^{x_H} g^2(x) \left[\frac{\ell}{f_K(x)} \right]^4 P_\Phi \left[\frac{\ell}{f_K(x)}, x \right] dx. \quad (7.90)$$

Such an expression is more general since it does assume the validity of the Poisson equation.

In practice, the field is smoothed with a filter, that is a window function $U(\theta, \theta_c)$, with angular radius θ_c as

$$\kappa(\theta_c) = \int \kappa(\theta') U(\theta', \theta_c) d^2\theta', \quad (7.92)$$

then the variance of the averaged field is given by

$$\langle \kappa^2 \rangle(\theta_c) = 2\pi \int_0^\infty d\ell \ell P_\kappa(\ell) \left[\int \theta U(\theta, \theta_c) J_0(\ell\theta) d\theta \right]^2, \quad (7.93)$$

where J_0 is a Bessel function (B.30).

7.1.4.3 Full-sky versus flat-sky derivations

The previous computation can be performed without making use of the flat-sky approximation. Starting from (7.80) rewritten as

$$\kappa(\mathbf{n}) = \frac{1}{2} \Delta_2 \tilde{\psi}(\mathbf{n})$$

where we recall that the deflecting potential integrated on the line of sight is

$$\tilde{\psi}(\mathbf{n}) = 2 \int_0^\chi \hat{g}(\chi) \Phi[\mathbf{x}(\mathbf{n}), \chi] d\chi \quad (7.94)$$

where Δ_2 is the Laplacian on the 2-sphere. As for any function on the sphere, it can be decomposed in spherical harmonics, in the same way as for the temperature anisotropies of the cosmic microwave background as

$$\tilde{\psi}(\mathbf{n}) = \sum_{\ell m} \psi_{\ell m} Y_{\ell m}(\mathbf{n}).$$

Decomposing the gravitational potential in Fourier modes in (7.94) and then the exponential with (B.21), we obtain

$$\tilde{\psi}(\mathbf{n}) = \sum_{\ell m} \left[8\pi i^\ell \int_0^\chi \hat{g}(\chi) j_\ell(k\chi) \Phi(k, \chi) Y_{\ell m}(\hat{k}) \frac{d^3 k}{(2\pi)^{3/2}} d\chi \right] Y_{\ell m}(\mathbf{n}),$$

where $\hat{g}(\chi) = g(\chi)/f_K(\chi)$, from which we deduce

$$\psi_{\ell m} = 8\pi i^\ell \int_0^\chi \hat{g}(\chi) j_\ell(k\chi) \Phi(k, \chi) Y_{\ell m}(\hat{k}) \frac{d^3 k}{(2\pi)^{3/2}} d\chi. \quad (7.95)$$

The power angular spectrum of $\tilde{\psi}$, defined as $\langle \psi_{\ell m} \psi_{\ell' m'}^* \rangle = \delta_{\ell\ell'} \delta_{mm'} C_\ell^{\psi\psi}$, is thus given by

$$C_\ell^{\psi\psi} = 16\pi \int_0^\chi \hat{g}(\chi) \hat{g}(\chi') j_\ell(k\chi) j_\ell(k'\chi') \mathcal{P}_\Phi(k, \chi, \chi') d\chi d\chi' \frac{dk}{k}, \quad (7.96)$$

where we recall that

$$\langle \Phi(k, \eta) \Phi^*(k', \eta') \rangle = \frac{2\pi^2}{k^3} \mathcal{P}_\Phi(k, \eta, \eta') \delta^{(3)}(k - k').$$

The convergence being given by $\kappa(n) = \frac{1}{2} \Delta_2 \psi(n)$, we deduce that

$$\kappa_{\ell m} = -\frac{\ell(\ell+1)}{2} \psi_{\ell m},$$

so that

$$C_\ell^{\kappa\kappa} = \frac{\ell^2(\ell+1)^2}{4} C_\ell^{\psi\psi}. \quad (7.97)$$

This expression has to be compared to the one we obtained previously in the flat-sky approximation using the Limber approximation that, rewritten in terms of $\tilde{\psi}$, takes the form

$$P_\psi(\ell) = 8\pi^2 \ell^{-3} \int_0^\chi \hat{g}^2(\chi) f_K(\chi) \mathcal{P}_\Phi \left[\frac{\ell}{f_K(\chi)}, \chi, \chi \right] d\chi$$

and

$$P_\kappa(\ell) = \frac{1}{4} \ell^4 P_\psi(\ell),$$

which is the same expression as in (7.90). On small angular scales (i.e. large ℓ) the two quantities $P_\kappa(\ell)$ and $C_\ell^{\kappa\kappa}$ coincide.

7.1.5 Cosmic shear

As will be seen, the statistical properties of the cosmic shear can be extracted from observations. We thus focus on the definition of these quantities and on their link with the convergence. It will be convenient to use a complex notation for the convergence, $\gamma = \gamma_1 + i\gamma_2$.

7.1.5.1 Kaiser and Squires algorithm

The shear and the convergence are defined from the second derivatives of the potential ψ and are thus not independent. An efficient method to reconstruct the projected mass distribution (the convergence) from the shear was proposed by Kaiser and Squires [36]. The expression (7.81) implies that

$$\gamma(\theta) = \frac{1}{\pi} \int K(\theta - \theta') \kappa(\theta') d^2 \theta', \quad (7.98)$$

where the function K , defined as

$$K = \frac{\theta_2^2 - \theta_1^2 - 2i\theta_1\theta_2}{\theta^4} = \frac{-1}{(\theta_1 - i\theta_2)^2}, \quad (7.99)$$

characterizes the shear induced by a point-like deflector. The convolution of the effective mass distribution $\Sigma \propto \kappa$ with this response will thus give the shear associated with this effective distribution. The aim is to invert this relation to reconstruct κ from γ .

For this, it is convenient to switch to Fourier space, where this convolution is converted into a simple multiplication

$$\hat{\gamma}(\ell) = \frac{1}{\pi} \hat{\mathcal{K}}(\ell) \hat{\kappa}(\ell), \quad (7.100)$$

for the Fourier modes, for all non-vanishing $\ell = (\ell_1, \ell_2)$. This can be rewritten as

$$\hat{\kappa}(\ell) = \frac{1}{\pi} \hat{\mathcal{K}}^*(\ell) \hat{\gamma}(\ell), \quad \hat{\mathcal{K}}(\ell) = \pi \frac{(\ell_1^2 - \ell_2^2 + 2i\ell_1\ell_2)}{\ell^2}. \quad (7.101)$$

The convergence, and thus the mass distribution, can thus be reconstructed from the cosmic shear as

$$\kappa(\theta) = \kappa_0 + \frac{1}{\pi} \int \mathcal{K}^*(\theta - \theta') \gamma(\theta') d^2\theta'. \quad (7.102)$$

The constant κ_0 is related to any uniform mass distribution that contributes to the convergence but not to the distortion. Notice that κ is real so that the imaginary part of the integral must vanish. What is obvious mathematically is no longer true with data since noise can generate such an imaginary part. Moreover, the integration can then no longer be performed on the entire plane but only on a support of the size of the receptor. The application of this method is thus not straightforward with real data where the noise, the need for smoothing, and the contribution from κ_0 , related to the mass-sheet degeneracy discussed in (7.44), must be taken into account.

An important consequence of these results is that [since (7.101) implies that $|\hat{\kappa}|^2 = \hat{\gamma} \cdot \hat{\gamma}^* = |\hat{\gamma}_1|^2 + |\hat{\gamma}_2|^2$] the convergence and the cosmic shear have the same power spectrum

$$P_\gamma(\ell) = P_\kappa(\ell). \quad (7.103)$$

7.1.5.2 Two-point statistics for the cosmic shear

The cosmic shear has two components that can be decomposed in various ways. Let us consider a spherically symmetric filter, centred in θ and of radius θ_c . The centre of the filter can be used to define the radial and orthoradial (also called tangent) components of the cosmic shear (see Fig. 7.12). Let φ be the unit vector that joins the centre of the filter to the point where the cosmic shear is defined (that is $\theta = \theta \exp i\varphi$). The shear is a spin-2 quantity (we recall that $\gamma(\theta) = \gamma \exp[2i\phi(\theta)]$). Thus, if we consider the local basis (e_r, e_t) at θ defined by $e_r = \exp i\varphi$ and $e_t = ie_r$, we can define in θ the two polarizations

$$\gamma_t(\theta) = -\text{Re} [\gamma(\theta) e^{-2i\varphi}], \quad \gamma_r(\theta) = -\text{Im} [\gamma(\theta) e^{-2i\varphi}], \quad (7.104)$$

exactly in the same way as for gravitational waves.

We can then define the different correlation functions between these components

$$\xi_r(\theta_c) = \langle \gamma_r \gamma_r \rangle, \quad \xi_t(\theta_c) = \langle \gamma_t \gamma_t \rangle. \quad (7.105)$$

They only depend on the size θ_c of the filter. Under a parity transformation, γ_t remains invariant while γ_r changes sign so that the correlation function $\langle \gamma_t \gamma_r \rangle = 0$. These two correlation functions can also be combined as

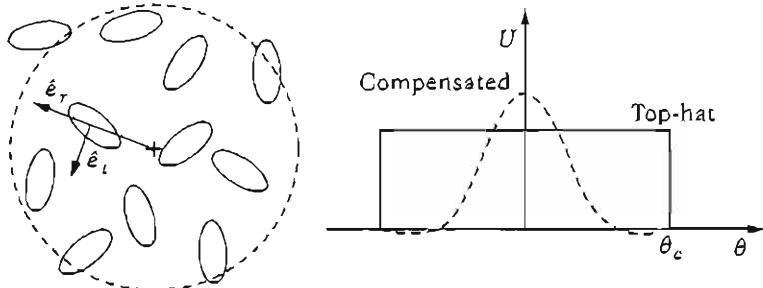


Fig. 7.12 (left): To compute the variance of the shear, the galaxy ellipticity is averaged with a filter of radius θ_c . Given the centre of the filter, the radial and orthoradial (tangent) components of the cosmic shear can be defined. (right): Two types of filters are usually used: top-hat and compensated filters.

$$\xi_{\pm}(\theta_c) = \xi_t \pm \xi_r. \quad (7.106)$$

Using these definitions before expanding $\kappa(\theta)$ into Fourier modes as (7.85) and using the relations (7.101) between the cosmic shear and the convergence, we obtain

$$\xi_+(\theta_c) = \int_0^\infty \frac{d\ell}{2\pi} \ell P_\kappa(\ell) J_0(\ell\theta_c) \quad \xi_-(\theta_c) = \int_0^\infty \frac{d\ell}{2\pi} \ell P_\kappa(\ell) J_4(\ell\theta_c). \quad (7.107)$$

Using (7.103), the average of the cosmic shear with a top-hat window function has the variance

$$\langle \gamma \cdot \gamma^* \rangle(\theta_c) = \int_0^\infty \frac{d\ell}{2\pi} \ell P_\kappa(\ell) \left[\frac{2J_1(\ell\theta_c)}{\ell\theta_c} \right]^2. \quad (7.108)$$

7.1.5.3 Aperture mass

For an arbitrary filter, the effect of an unknown mass sheet (7.102) biases the reconstruction. However, if a compensated filter is used, i.e. one satisfying

$$\int \theta U(\theta, \theta_c) d\theta = 0,$$

then any constant contribution to the integration (7.92) would play no role. Defining the aperture mass by

$$M_{ap} \equiv \int d^2\theta U(\theta, \theta_c) \kappa(\theta), \quad (7.109)$$

one can then show that this quantity can be expressed in terms of the tangent shear alone as

$$M_{ap} = \int d^2\theta Q(\theta, \theta_c) \gamma_t(\theta), \quad Q(\theta, \theta_c) = \frac{2}{\theta^2} \int d\theta' \theta' U(\theta', \theta_c) - U(\theta, \theta_c). \quad (7.110)$$

A very popular and useful family of windows [11] is

$$U(\theta, \theta_c) = \frac{3}{\pi \theta_c^2} \left[1 - \left(\frac{\theta}{\theta_c} \right)^2 \right] \left[1 - 3 \left(\frac{\theta}{\theta_c} \right)^2 \right] \Theta(\theta < \theta_c). \quad (7.111)$$

This filter is a bandpass filter of vanishing average, which gives the variance

$$\langle M_{ap}^2 \rangle = \frac{288}{\pi} \int d\ell \ell P_\kappa(\ell) \left[\frac{J_4(\ell \theta_c)}{\ell^2 \theta_c^2} \right]^2. \quad (7.112)$$

7.1.5.4 Relations between the power spectra

All of these two-point correlation functions are related to the convergence power spectrum. So they are not independent. For instance, one can show, using the properties of the Bessel functions, that

$$P_\kappa(\ell) = 2\pi \int_0^\infty d\theta_c \theta_c J_0(\ell \theta_c) \xi_+(\theta_c) = 2\pi \int_0^\infty d\theta_c \theta_c J_4(\ell \theta_c) \xi_-(\theta_c). \quad (7.113)$$

Similarly $\langle M_{ap}^2 \rangle$ and $\langle \gamma \gamma^* \rangle$ can also be expressed as a function of ξ_\pm [37].

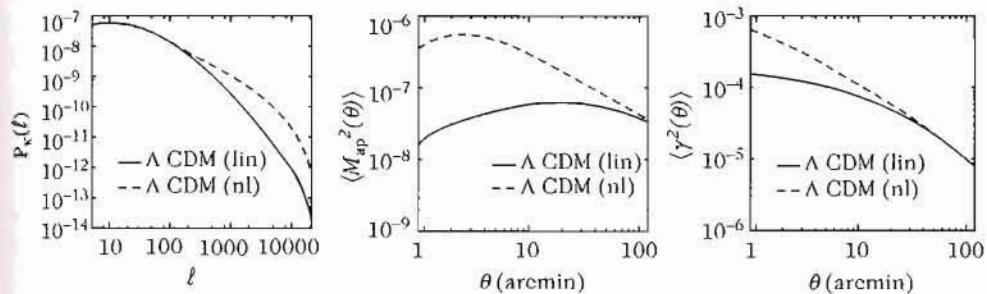


Fig. 7.13 Prediction of the different observables for a Λ CDM model with the contribution from the linear (solid lines) and non-linear (dashed lines) regimes of the power spectrum (see Chapter 5, notice that even in this regime lensing can be treated in a linear way). From left to right, the convergence power spectrum, $P_\kappa(\ell)$, the aperture mass $M_{ap}^2(\theta)$, and the cosmic shear top-hat variance, $\langle \gamma^2(\theta) \rangle$. From Ref. [34].

7.1.6 Measurement of the cosmic shear

The key in the measurement of the cosmic shear relies on the measurement of the shape of background galaxies and on their ellipticity (see Refs. [11, 12, 38] for details and problems). If the image of such a galaxy is represented by an ellipticity $\varepsilon = \varepsilon_1 + i\varepsilon_2 = (1-r)/(1+r) \exp(2i\phi)$, $r = b/a$ being the ratio between the major and minor axes, then

$$\langle \varepsilon \rangle = \left\langle \frac{\gamma}{1 - \kappa} \right\rangle.$$

In the regime of weak distortion ($\kappa \ll 1$), the ellipticities give access to the shear (see Fig. 7.14).

The measurement is thus based on the crucial assumption that the source galaxies have a random and uncorrelated intrinsic ellipticity so that any deviation from a random distribution and any correlation in the ellipticities are attributed to lensing effects.

Assuming that the ellipticities are distributed with a dispersion $\sigma_e \sim 0.3$, as suggested from surveys of close galaxies in the Hubble Deep Field, then one can measure an ellipticity, ε , induced by a gravitational shear of the order of $\gamma \sim 0.1$ by averaging over the number of galaxies

$$N > \left(\frac{\sigma_e}{\varepsilon}\right)^2 \simeq 10.$$

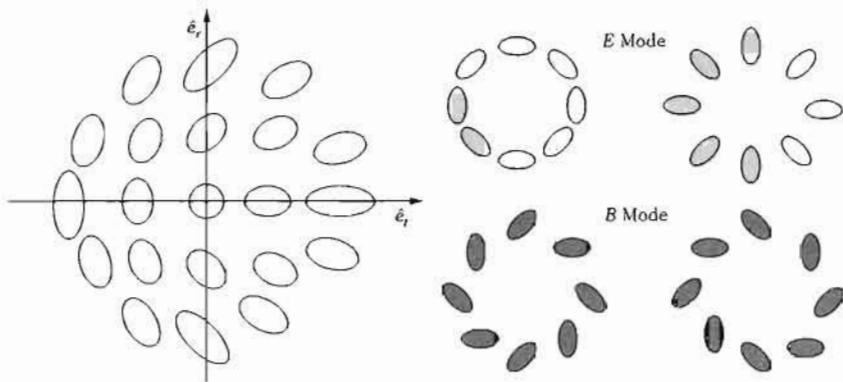


Fig. 7.14 (left): Value of (γ_r, γ_t) as a function of the shape of the galaxy with respect to the local reference frame. The ellipticity is invariant under a rotation of angle π . (right): E and B modes. Only E modes are produced by a gravitational field.

It therefore seems easy to measure a gravitational shear of 10%. Nevertheless, measuring a shear of the order of 1% requires 900 galaxies, which is hardly achievable on small angular scales, unless we average over many fields of 1 arcmin.

Nevertheless, many observational problems exist (see Refs. [12, 38] for detailed discussions on the observational aspects). In particular, one should resolve problems (1) from the degeneracy with a mass sheet (for instance, the statistic $\langle M_{ap}^2 \rangle$ avoids it), (2) from the angular resolution that is limited by the surface density of galaxies, (3) from the redshift distribution of the sources that is important for calibration, (4) from intrinsic alignments that can bias the measurements, (5) from the atmospheric turbulence that makes every object more circular and thus dilutes the signal and (6) from the optical aberrations that introduce artificial ellipticities. Relations such as (7.113) can allow us to control the level of these sources of error for which the effects on the shear do not a priori satisfy these constraints.

By analogy to the terminology used for the polarization of the cosmic microwave background, one can define E and B modes as

$$\begin{aligned} \Delta E &= (\partial_1^2 - \partial_2^2)\gamma_1 + 2\partial_1\partial_2\gamma_2, \\ \Delta B &= -2\partial_1\partial_2\gamma_1 + (\partial_1^2 - \partial_2^2)\gamma_2. \end{aligned} \quad (7.114)$$

As can be seen from the definitions of the shear, gravitational lensing induces only E modes (see Fig. 7.14) and the E polarization is in fact simply the convergence. The two components of the shear are actually not independent, which implies, for instance, that the imaginary part of (7.102) vanishes. Since the definitions of E and B are non-local, their reconstruction in real space depends, at any point, on the value of the shear in the whole space. A way to solve this problem is to filter the shear by a filter that eliminates long wavelengths. This can be achieved by considering a compensated filter. In particular, one can show that $\langle M_{\text{ap}}^2 \rangle$ is only sensitive to the E modes while

$$M_{\perp}(\theta_c) = \int d^2\theta Q(\theta, \theta_c) \gamma_r(\theta), \quad (7.115)$$

is only sensitive to the B modes. M_{\perp} must strictly vanish if the distortion has a gravitational origin so that any deviation of M_{\perp} from zero can be used to control the data analysis accuracy.

7.1.6.1 Observational status

The first detection of the cosmic shear was performed in March 2000 [39] and then by three other groups [40–42]. These surveys counted around 10^5 images of galaxies on a surface of around 1 deg^2 . Figure 7.15 summarizes the results of these various analyses for the variance of the shear and the various two-point correlators.

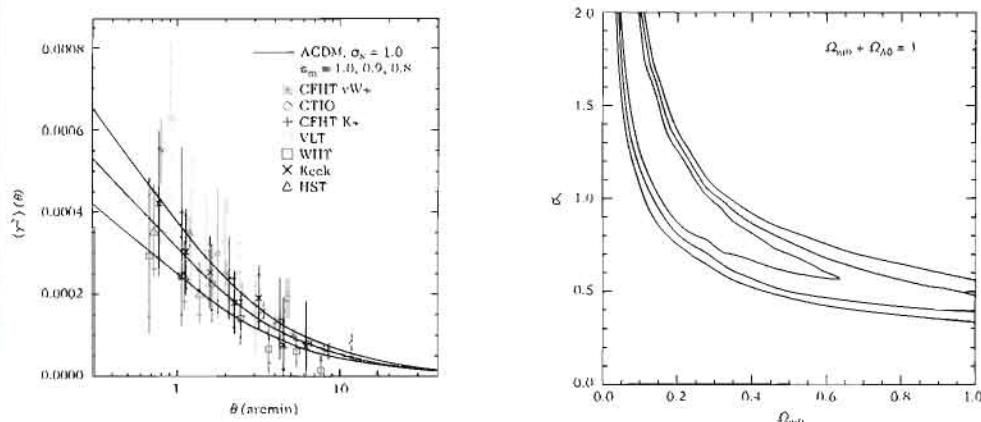


Fig. 7.15 (left): Compilation of the first observations of the cosmic shear variance (using a top-hat filter). The three curves correspond to the three redshifts 1, 0.9 and 0.8, and data are in good agreement with the behaviour of (7.116). (right): Constraints for the parameters Ω_{m0} and σ_8 obtained from the survey VIRMOS-DECART [43]. The contours are at 1, 2 and 3σ and they were obtained after marginalization over z_s and n .

These measurements allow us to constrain the cosmological parameters. The cosmic shear is especially sensitive to the total matter density and to the normalization of the power spectrum. For instance, one can show that for Gaussian perturbations with spectral index n ,

$$\langle \kappa^2 \rangle^{1/2}(\theta) \simeq 0.01 \sigma_8 \Omega_{m0}^{0.75} z_s^{0.8} \left(\frac{\theta}{1^\circ} \right)^{-(n+2)/2}, \quad (7.116)$$

for sources at redshift z_s . The non-Gaussianity induced by the non-linear dynamics (see Chapter 5) has a different dependence

$$\frac{\langle \kappa^3 \rangle}{\langle \kappa^2 \rangle^2}(\theta) \simeq -42 \sigma_8 \Omega_{m0}^{-0.8} z_s^{-1.35}. \quad (7.117)$$

This allows us to obtain strong constraints on the parameters Ω_{m0} and σ_8 (see Fig. 7.15). The latest CFHTLS data [44] span 57 square degrees and allow us to measure the 2-point shear statistics from 1 arcmin to 4 degrees. Its analysis leads to the constraint

$$\sigma_8 \left(\frac{\Omega_{m0}}{0.25} \right)^{0.64} = 0.785 \pm 0.043,$$

assuming $K = 0$. Assuming a mean redshift of 0.5, the largest physical scale probed by this analysis is 85 Mpc. Using only data with $\theta > 85$ arcmin, so that only scales in the linear regime are taken into account, one deduces that

$$\sigma_8 \left(\frac{\Omega_{m0}}{0.25} \right)^{0.53} = 0.771 \pm 0.029.$$

This illustrates the potential of future wide-field observations to probe the large-scale structure.

7.1.7 Lensing on the cosmic microwave background

After decoupling, photons from the cosmic microwave background are decoupled from matter and propagate along geodesics. As seen several times in this chapter, these geodesics are deviated by the gravitational field of the large-scale structure. The observed CMB temperature and polarization fields $\hat{T} = (\hat{\Theta}, \hat{Q} \pm i\hat{U})$ in a given direction n are related to their values at the time of decoupling by

$$\hat{T}[n] = T[n + \xi(n)], \quad (7.118)$$

where ξ is the displacement field induced by gravitational lensing. For small displacements, each quantity can be expanded to second order as

$$\hat{X}(n) = X(n) + \xi^i(n)X_i + \frac{1}{2}\xi^i(n)\xi^j(n)X_{ij}. \quad (7.119)$$

7.1.7.1 Effects on the temperature

The impact of the gravitational lensing on the angular power spectrum of the temperature anisotropies can be illustrated in the flat-sky approximation, in which we can

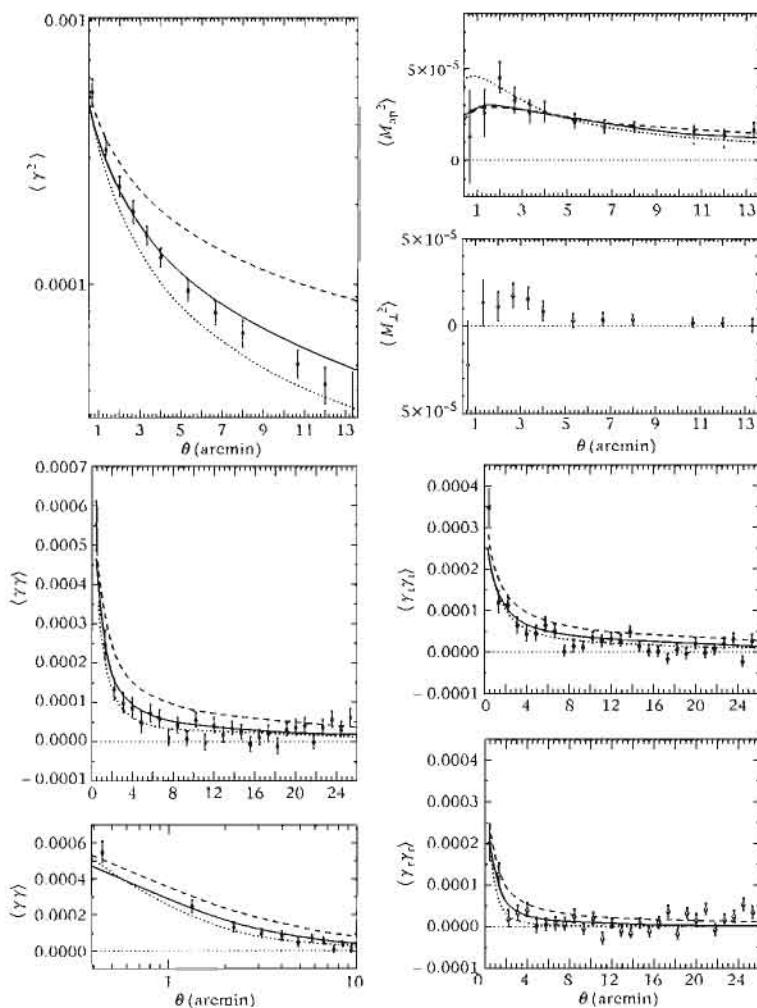


Fig. 7.16 Measurements of two-point statistics in the survey VIRMOS-DESCART [43]. (top left): Variance of the shear with a top-hat filter, (top right) aperture mass of the E and B modes. The B modes must a priori vanish so that the deviation of M_{ap}^2 with respect to zero can control the data analysis, and in particular the systematic effects. (bottom left): Correlation function of the shear and (bottom right) tangent and radial projections of the shear.

expand the cosmic microwave background temperature anisotropies in Fourier modes as

$$\Theta(\mathbf{n}) = \int \frac{d^2\ell}{2\pi} \Theta(\ell) e^{i\ell \cdot \mathbf{n}}. \quad (7.120)$$

For a Gaussian field, the two-point correlation function is given by

$$C(\ell) = \langle \Theta(0)\Theta(n) \rangle = \int \frac{d^2\ell}{(2\pi)^2} C_\ell e^{i\ell \cdot n}, \quad (7.121)$$

with $\langle \Theta(\ell)\Theta(\ell') \rangle = C_\ell \delta^{(2)}(\ell + \ell')$. According to (7.118), this angular correlation function is modified by lensing effects as

$$\langle \hat{\Theta}(0)\hat{\Theta}(n) \rangle = \langle \hat{\Theta}(0 + \xi(0))\hat{\Theta}(n + \xi(n)) \rangle \quad (7.122)$$

$$\begin{aligned} &= C(\theta) + \langle \xi^i(0)\xi^j(n) \rangle \langle \Theta_i(0)\Theta_j(n) \rangle + \frac{1}{2} \langle \xi^i(0)\xi^j(0) \rangle \langle \Theta_{ij}(0)\Theta(n) \rangle \\ &\quad + \frac{1}{2} \langle \xi^i(n)\xi^j(n) \rangle \langle \Theta_{ij}(n)\Theta(0) \rangle. \end{aligned} \quad (7.123)$$

This function has to be expanded to second order in the perturbation decomposition in ξ to obtain the leading contribution to C_ℓ . Indeed, since the average over the displacement vanishes, $\langle \xi^i \rangle = 0$, it is at this order that the contributions from the lensing coupling in the two-point correlation functions appear.

Using the relation between ξ and the gravitational potential and then performing a Fourier transform, we obtain

$$\hat{C}_\ell = C_\ell \left[1 - \int \frac{d^2\ell'}{(2\pi)^2} \frac{(\ell' \cdot \ell)^2}{\ell'^4} \bar{P}(\ell') \right] + \int \frac{d^2\ell'}{(2\pi)^2} \frac{(\ell' \cdot \ell)^2 - \ell'^4}{\ell'^4} \bar{P}(\ell') C_{|\ell-\ell'|}, \quad (7.124)$$

where \bar{P} is the power spectrum integrated along the line of sight.

Two effects appear, (1) a renormalization that affects the entire spectrum and that is not directly detectable and (2) a mode coupling. The second term in the previous expression is a convolution that tends to smooth out the angular power spectrum of the temperature anisotropies. This term is too weak to affect the first acoustic peak so that its effect will mainly affect the tail of the spectrum. This implies that the angular power spectrum of the temperature anisotropies, by itself, only gives little information on lensing.

There are other effects on the anisotropies, for instance, morphologic effects. In particular, the peaks and troughs of the temperature map are modified. These effects are related to the departures from the Gaussian properties of the perturbations. For instance, the leading term of the four-point function of the temperature field is

$$\langle \hat{\Theta}(n_1)\hat{\Theta}(n_2)\hat{\Theta}(n_3)\hat{\Theta}(n_4) \rangle_c = \langle \xi^i(n_1)\xi^j(n_3) \rangle \langle \partial_i \hat{\Theta}(n_1)\hat{\Theta}(n_2) \rangle \langle \partial_j \hat{\Theta}(n_3)\hat{\Theta}(n_4) \rangle, \quad (7.125)$$

plus its permutations.

7.1.7.2 Effects on polarization

The Stokes parameters are modified according to (7.119). Only the relation between the observation and the emission points change and gravitational lensing induces no rotation of the polarization vector ($Q \pm iU$). Moreover, no polarization can be created, since if the polarization field vanishes initially, then it will remain so.

However, gravitational lensing will change the properties of the polarization field. For this, we consider the decomposition of this field into E and B modes, (6.204).

Furthermore, one should establish the analogue of the relation (7.119) but for the second derivatives

$$\partial_i \partial_j [\hat{X}(\mathbf{n})] = [X_{,ij} + 2\xi_{(i}^l X_{,j)l} + \xi_i^l X_{,jl}] (\mathbf{n}), \quad (7.126)$$

at first order in ξ . This relation can then be used to obtain the effects on the E and B modes,

$$\Delta \hat{E} = (1 - 2\kappa) \Delta E + \xi \cdot \nabla (\Delta E) - 2\delta^{ij} (\gamma_i \Delta P_j + \nabla \gamma_i \cdot \nabla P_j), \quad (7.127)$$

$$\Delta \hat{B} = (1 - 2\kappa) \Delta B + \xi \cdot \nabla (\Delta B) - 2\varepsilon^{ij} (\gamma_i \Delta P_j + \nabla \gamma_i \cdot \nabla P_j), \quad (7.128)$$

where $P = Q + iU$. The lensing effects can thus be decomposed as:

- A displacement, given by the term $(1 + \xi \cdot \nabla) \Delta(E, B)$, which is the perturbative expansion of the Laplacian of (E, B) at the point $\mathbf{n} + \xi$. This term is identical to that appearing for the temperature.
- A magnification, controlled by the term -2κ that introduces the convergence. Such a term is not surprising.
- A mixture of E and B modes, dictated by the contraction between the cosmic shear and the polarization vector. This coupling term is composed of two contributions that will dominate at different scales, the one involving the gradient of the shear dominating at small scales.

More details on the effects of lensing on temperature and polarization are given in Refs. [45, 46].

As seen in Chapter 6, the B modes are only generated by primordial gravitational waves. Assuming that initially there is only a scalar polarization, then the gravitational lensing induces B modes,

$$\Delta \hat{B} = -2\varepsilon^{ij} (\gamma_i \Delta P_j + \nabla \gamma_i \cdot \nabla P_j). \quad (7.129)$$

This has an important consequence: the detection of B modes will bring much information on the gravitational lensing by large-scale structure but, moreover, it will tend to hide the B modes generated by the primordial gravitational waves. Figure 7.17 illustrates these effects on the power spectrum.

7.2 Evidence for the existence of dark matter

In the modern cosmological model, dark matter plays a central role in the formation of large-scale structures. We review the different ‘proofs’ of the existence, or more exactly of the necessity, of this matter. These conclusions rely on the validity of general relativity to describe gravity, which we will assume here. To finish, we will also consider the possibility of a modification of this law of gravity.

7.2.1 Dark matter in galaxies

The most convincing and direct proof of the existence of dark matter comes from the observation of the galaxy rotation curves in spiral galaxies. Spiral galaxies have a flat disk in which stars follow circular orbits. The rotation curves give the orbital velocity of stars as a function of their distance to the centre of the galaxy.

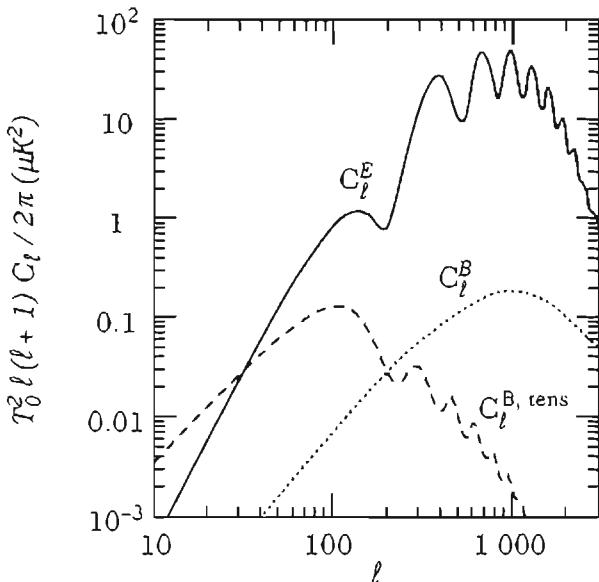


Fig. 7.17 *B* mode of the cosmic microwave background polarization induced by gravitational lensing (dotted line) compared to that induced by primordial gravitational waves (dashed line) and to *E* modes (solid line) assuming that $T/S = 1$. From Ref. [46].

7.2.1.1 Rotation curves and dynamical mass

Newtonian dynamics implies that the Keplerian velocity depends on the mass contained inside the orbit

$$\frac{v^2(r)}{r} = \frac{G_N M(< r)}{r^2}, \quad (7.130)$$

where $M(< r) = 4\pi \int \rho(r) r^2 dr$. The surface brightness of spiral galaxies is well fitted by a function that decreases exponentially as

$$I(r) = I_0 \exp\left(-\frac{r}{R_d}\right), \quad (7.131)$$

where the disk diameter, R_d , is a characteristic length of the order of a few kpc (4 kpc for our Galaxy, 6 kpc for M31, for instance). So, if the stars were the major contributors to the galaxy mass, i.e. if there were no non-luminous matter, then $M(< r)$ would become constant beyond the optical disk. As a result,

$$M(< r) \rightarrow \text{const.}, \quad v(r) \propto \frac{1}{\sqrt{r}}, \quad r \gtrsim R_d. \quad (7.132)$$

Actually, the rotation curves, which are particularly well measured for spiral galaxies, do not reproduce this behaviour. The rotation curves tend towards a non-vanishing asymptotic value v_∞ and the Keplerian regime never seems to be reached. If $v \rightarrow v_\infty$,

$M(<r) \propto r$ at large distances, which indicates the existence of a non-luminous matter with density profile behaving as $\rho \propto 1/r^2$ at large distances from the centre.

At large distances, we can express this asymptotic velocity, v_∞ as

$$v_\infty \equiv \sqrt{G_N Ma_0}, \quad (7.133)$$

where a_0 is a constant with dimensions of an acceleration. This implies in particular that at large distances the deflection angle, see (7.1), is given by

$$\alpha \equiv \frac{2\pi}{c^2} \sqrt{G_N Ma_0}. \quad (7.134)$$

7.2.1.2 Description of the rotation curves

To better understand the implications of the rotation curves, we need to introduce some quantities used in the astrophysical literature (see Ref. [47] for a review).

The dynamical mass is the mass deduced from the observation of these rotation curves

$$G_N M_{\text{dyn}}(r) = rv^2(r). \quad (7.135)$$

The measurement of the dynamical mass relies on the measurement of the stars rotation velocities in the exterior zones of the disk from their Doppler effect that gives access to the line of sight velocity

$$v_r = v_{\text{gal}} + v \sin i, \quad (7.136)$$

where i is the inclination of the galaxy disk and v_{gal} is the galaxy velocity. We get the radius from the last observed point, r_{\max} , at which the velocity is called v_{\max} or v_∞ .

However, not all of the dynamical mass is in the form of dark matter. Observations of the CO molecular lines allow us to reconstruct the internal part of the rotation curve and the gas contribution in this zone where it dominates. Radio observations from the HI gas, using the 21 cm hydrogen line, allow us to reconstruct the most exterior part, beyond the stellar disk, where dark matter dominates.

We can thus decompose the rotation curve into a contribution from the stellar disk and the gas as

$$v^2 = v_{\text{baryon}}^2 + v_{\text{DM}}^2, \quad (7.137)$$

where the baryonic component is itself decomposed into

$$v_{\text{baryon}}^2 = v_{\text{gas}}^2 + v_{\text{stars}}^2.$$

Each of these velocities is related to the mass contribution of the corresponding component, where the link is provided by the Kepler law

$$v_X^2 \equiv \frac{G_N M_X}{r}.$$

Given the mass-luminosity ratio m_*/L_* of the stellar population of the galaxy, one can reconstruct the mass profile of the halo (see Fig. 7.18) by distinguishing between

the contributions from the gas, the stars and, by taking the difference, from the dark-matter halo. In fact, from the luminosity profile (7.131), one can define, for spiral galaxies, a surface density profile by

$$\Sigma = \Sigma_0 \exp\left(-\frac{r}{R_d}\right). \quad (7.138)$$

The optical radius, R_{opt} , is defined as the radius containing 80% of the luminous energy. It is related to the disk radius by $R_{\text{opt}} \simeq 3.2 R_d$. We obtain equivalent profiles for the gas.

In conclusion, we see that we have two important notions of mass: the dynamical mass, related to the rotation curves, and the baryonic mass, assumed to be proportional to the luminous mass. We have a third notion of mass to consider, which is the mass M_{lens} that can be determined by lensing. In the standard dark matter model, these masses should satisfy

$$M_{\text{baryon}} < M_{\text{DM}} \simeq M_{\text{dyn}} \simeq M_{\text{lens}}. \quad (7.139)$$

In what follows, we shall test this relation against observations.

7.2.1.3 Spiral galaxies

Milky Way

The rotation curve of the Milky Way (MW) is currently known with a great precision. In the central part of our Galaxy, the stars velocity dispersion is explained by the presence of a black hole of mass $M_\bullet \simeq (2.87 \pm 0.15) \times 10^6 M_\odot$ that was determined from the orbital parameters of stars orbiting close to the galactic centre [48]. The presence of supermassive black holes is today accepted in numerous galaxies.

Our Galaxy can be roughly decomposed as (a) a high-density core, containing a black hole, (b) a central region ($\lesssim 100$ pc) where the density rapidly increases before (c) it reaches a maximum at around 300 pc and then decreases towards a minimum at around 2 kpc before (d) a gradual pull-up to reach the disk maximum at around 6 kpc and (e) a plateau with holes at around 8 kpc and 15 kpc.

Taking as a reference the Sun, the dynamical mass (7.135) is then

$$M(< r) = 9.6 \times 10^{10} M_\odot \left(\frac{v}{220 \text{ km} \cdot \text{s}^{-1}} \right)^2 \left(\frac{r}{100 \text{ kpc}} \right). \quad (7.140)$$

Since the luminosity of the MW is estimated as $L_{\text{MW}} = 2.3 \times 10^{10} L_\odot$, we obtain the mass/luminosity ratio, Q , of the Milky Way

$$Q_{\text{MW}} \simeq 50 Q_\odot \left(\frac{R_{\text{halo}}}{100 \text{ kpc}} \right). \quad (7.141)$$

The last detectable gas is found around 20 kpc, for which there is still no sign of any Keplerian regime, so that $R_{\text{halo}} > 20$ kpc. The dynamics of globular clusters and of the Magellan clouds allows us to deduce that $R_{\text{halo}} > 75$ kpc, so that

$$Q_{\text{MW}} \simeq (10 - 40) Q_{\odot}. \quad (7.142)$$

More locally, the comparison between the mass of the close stars and gas to the dynamical mass allows us to obtain that

$$Q_{\text{MW}} \simeq 5 Q_{\odot} \quad (7.143)$$

in a region of around a hundred parsecs. This limit, obtained from the dynamics of stars normal to the plane of the disk, is called the Oort limit.

The Local Group, i.e. mainly the Andromeda galaxy and a few satellite galaxies, can be considered as an isolated system since the closest galaxy, M81, is at a distance of around 3 Mpc. Andromeda and the Milky Way are separated by around $r = 10^3$ kpc and have a relative velocity of $v \sim 10^2 \text{ km} \cdot \text{s}^{-1}$, so that $vH_0/rc \sim \mathcal{O}(1)$, which shows that the system is gravitationally bound. This means that the dynamical mass is of the order of $10^{12} M_{\odot}$. A more precise estimate gives $M = (3.2 - 5.5) \times 10^{12} M_{\odot}$ and the total luminosity of both companions can be estimated to $L \sim 4.2 \times 10^{10} L_{\odot}$ so that

$$Q_{\text{group}} = (76 - 130) Q_{\odot}. \quad (7.144)$$

Spiral galaxies

Today, the rotation curves of several hundreds of spiral galaxies have been catalogued. They behave differently in the internal zone, depending mainly on the importance of the bulge. Depending of the type of galaxies, these rotation curves have different morphologies but all of them seem to reach a plateau.

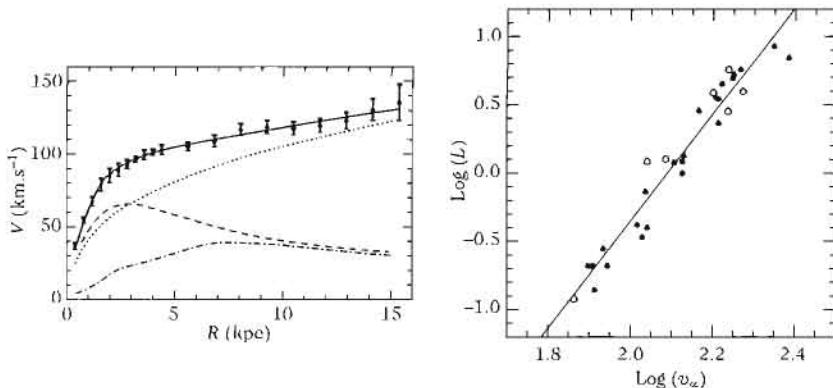


Fig. 7.18 (left): Rotation curve of the galaxy M33. The observations (points and circles) are compared to a model of dark-matter halo (solid line). The contributions from the halo (dashed-dotted line), from the stellar disk (short dashes) and from the gas (long dashes) are also represented. (right): Tully–Fisher relation for various spiral galaxies.

Figure 7.18 represents the rotation curve of the galaxy M33 and many other examples can be obtained, see e.g. Ref. [49]. Notice the fact that the rotation curve remains

flat while the dominant matter changes from stars and gas to dark matter; this may look like a conspiracy. It seems that a ‘tuning’ exists between the baryonic matter and the dark matter, without which the plateau could have had some bumps or holes. It seems that the dissipative evolution of the baryons in the potential wells of dark matter is such that both profiles adjust to each other.

There exists a relation, without too much dispersion, between the luminosity of a spiral galaxy and the maximal velocity, v_∞ of its rotation curve plateau. This Tully–Fisher relation

$$L \propto v_\infty^\alpha, \quad \alpha = 3.9 \pm 0.2, \quad (7.145)$$

tells us that for these galaxies, there exists a scaling relation between their mass and size. This relation is no longer valid for galaxies of small luminosity for which the mass of the gas must be taken into account in the luminosity.

Variations around the Tully–Fischer relation

The Tully–Fisher law (7.145) relates the total luminosity to the velocity v_∞ , for spiral galaxies. A generalized or baryonic Tully–Fischer law has also been proposed [50]. This law relates the total baryonic mass to v_∞ as

$$M_{\text{baryon}}(r < r_{\max}) \propto v_\infty^\alpha, \quad (7.146)$$

with $\alpha \sim 4$. This law therefore includes the contribution of both the gas and the stars and thus becomes general for any kind of spiral galaxy (dwarf, low surface brightness, etc.).

It has been conjectured [51] that dark matter follows the gas distribution and, in particular, that their rotation curves were homothetic

$$v_{\text{DM}}^2 = \gamma v_{\text{gas}}^2. \quad (7.147)$$

With this hypothesis, the total rotation curve can be decomposed into two contributions, that of the stars and that of the gas, and only introduces one arbitrary parameter $v^2 = v_{\text{stars}}^2 + (1 + \gamma)v_{\text{gas}}^2$. Various studies [52, 53] have shown that this hypothesis seems in to be in good agreement with observations and that the parameter γ must be in the range $\gamma = 3 - 16$.

Finally, at the characteristic scales of these galaxies, the ratio between the surface densities of dark and baryonic matter is of the order of [52]

$$\frac{\Sigma_{\text{DM}}}{\Sigma_{\text{baryon}}} = 8.5 \pm 1.5. \quad (7.148)$$

These relations have led to the hypothesis that the dark matter from the galaxy halos could be in the form of a gas too cold to be detectable [54].

Universal profiles

The description of the galaxy halos was initially addressed by establishing empirical density profiles [55] that have the property of decreasing as r^{-2} at large distances. An example of these profiles is the isothermal model [see (7.38) for its derivation]. This

model is singular at the centre and has been improved with an isothermal model with core radius.

A commonly used empirical parameterization describes the halo with the profile

$$\rho_{\text{DM}} = \rho_s \frac{1 + 3(r/a)^2}{3[1 + (r/a)^2]^2}. \quad (7.149)$$

Its only noticeable difference from the isothermal model with core radius is its behaviour in the few central kpc. It seems that all of the halos of spiral galaxies follow a law of this kind [55]. This suggests searching for a general dark-matter halo profile of the form

$$\rho_{\text{DM}}(a) = \rho_0(a) f\left(\frac{r}{a}\right), \quad (7.150)$$

f being a universal function. Let us stress that the function ρ_0 could depend on other parameters such as the galaxy luminosity. The dark-matter mass is thus given by

$$M_{\text{DM}}(r) = 4\pi\rho_0(a)a^3 F\left(\frac{r}{a}\right), \quad F(x) = \int_0^x u^2 f(u) du. \quad (7.151)$$

Observations [56] tend to prove that, under the assumption of a profile of the form $\rho \propto 1/(r^2 + r_c^2)$, the dynamical mass $M_{\text{dyn}}(a) \propto a^2$ and $a \sim 2.3R_d$, which implies in particular that almost all of the baryonic component of the galaxy is in the core of the dark-matter halo since it mainly develops from r_c .

Moreover, in spiral galaxies the baryonic matter is distributed in the form of a disk so that $M_{\text{baryon}}(a) \propto R_d^2 \propto a^2$, from which we can deduce that $M_{\text{DM}}(a) \propto a^2$. Using the form (7.151), this implies that

$$\rho_0(a) \propto a^{-1} \implies \rho_0(a) \equiv \rho_{\text{crit}} \frac{\ell_0}{a}. \quad (7.152)$$

Furthermore, from these observations one understands that not only do we have a generalized Tully-Fischer law, but in fact it is true for each component since $M_{\text{DM}} \propto a^2$, so that $v_{\text{DM}}^2(u) \propto M_{\text{DM}}(a)/a \propto a$ and thus $v_{\text{DM}}^4(a) \propto a^2 \propto M_{\text{DM}}(a)$.

As will be seen from numerical simulations, every profile mainly differs by its behaviour at the centre and at very large distances. It is difficult to determine from the kinetics of the central region whether the distribution of dark matter is peaked or if the galaxy has a core. The dynamics of this region is indeed in general dominated by visible matter. When this is no longer the case, as for galaxies with low surface brightness, the central region is so small that the observational resolution becomes a limiting factor.

7.2.1.4 Other observations

Elliptical galaxies

Elliptical galaxies are particularly interesting for the study of dark matter as they have a simpler geometry and their baryonic mass is almost uniquely in the form of stars (they contain a small amount of gas). Information on these galaxies comes from

three sources: optical observations, observations concerning dark matter (gravitational lensing, virial, rotation curves) and numerical simulations.

Optical observations have shown that there exists a relation between the luminosity L_* and the central velocity dispersion, σ_0 ,

$$L_* \sim \sigma_0^4. \quad (7.153)$$

This *Faber–Jackson relation* is the analogue of the *Tully–Fischer relation*. Thousands of galaxies from SDSS [57] have been used to establish that

$$\sigma_0 \sim L_*^{0.25 \pm 0.012}. \quad (7.154)$$

Moreover, these optical observations were able to establish that there was a very precise relation [58] between the effective radius, R_{eff} , i.e. the radius containing half of the luminosity, the central velocity dispersion and the luminosity L_* . The SDSS [57] has established that

$$R_{\text{eff}} = \sigma_0^{1.49 \pm 0.05} \mu^{-0.75 \pm 0.01}, \quad (7.155)$$

with $\mu \equiv L_*/R_{\text{eff}}^2$. In logarithmic coordinates, this relation defines the fundamental plane of the galaxy. The projection of this relation implies that $R_{\text{eff}} \sim L_*^{0.63 \pm 0.025}$. It also leads to the conclusion that the surface density $\Sigma_*(r)$ can be described by a function of three parameters.

It has indeed been shown observationally that a profile with three parameters, called the *Sersic profile*, of the form

$$\ln \left(\frac{\Sigma_*}{\Sigma_{\text{eff}}} \right) = -b \left(X^{1/n} - 1 \right), \quad (7.156)$$

with $\Sigma_{\text{eff}} = \Sigma(R_{\text{eff}})$ and $X = R/R_{\text{eff}}$, reproduces well the observations. In particular, for any elliptical galaxy, $b(n) = (2n - 0.324)$ with n between 0.1 and 100. The dispersion of n is today understood as arising partly from the triaxiality of the elliptical galaxies.

The observational data lead to the conclusion that stars dominate at the centre. In particular, the dynamical mass varies very little until $2R_{\text{eff}}$ which shows that dark matter is either subdominant at the centre or that it has a noticeably constant density or finally that its density follows that of the light distribution. Observations of the stellar velocity distribution tend to prove that the first hypothesis is favoured. However, we should recall that elliptical galaxies are formed by successive fusion with early galaxies that induces the existence of radial stellar orbits, causing the dynamical analysis of the stars to be more delicate. At large radius, the observation of multiple images of quasars shows that dark matter dominates the matter distribution. Despite the simple geometry of these galaxies, the distribution of dark matter remains poorly known, by lack of precise measurements of the stellar dynamics.

Low surface brightness galaxies

Low surface brightness galaxies seem to be completely dominated by dark matter, with the stellar population contributing only marginally to the total mass. In particular, the dark-matter density profile does not appear to be peaked at the centre.

The relative contributions and distributions of dark matter in various kinds of galaxies contain important information for understanding the relation between dark matter and ordinary matter. Elliptical and spiral galaxies with low surface brightness give access to limiting cases, the consequences of which have not yet been drawn, for lack of precise enough observations.

Relation with the mass of the central black hole

Just like the Milky Way, it seems that elliptical galaxies and the bulges of most of the spiral galaxies contain a central black hole. A recent empirical law has been established between the mass of the central black hole, M_\bullet , and the velocity dispersion of the bulge, σ_v ,

$$\ln\left(\frac{M_\bullet}{M_\odot}\right) = (4.02 \pm 0.32) \ln\left(\frac{\sigma_v}{200 \text{ km} \cdot \text{s}^{-1}}\right) + (8.13 \pm 0.06). \quad (7.157)$$

This relation has been established from the observation of 30 spiral galaxies for which the mass of the central black hole and the rotation curve could be measured.

Since the velocity dispersion is proportional to $L_*^{1/4}$ (Faber–Jackson for ellipticals or Tully–Fisher for spirals) we expect that

$$L_* \sim M_\bullet.$$

The observations seem to indicate effectively that

$$M_\bullet \sim 10^{-3} M_*,$$

M_* being the stellar mass of the elliptical galaxy or of the bulge of the spiral galaxy.

This relation ($M_\bullet - \sigma_v$) brings information that can be useful for the understanding of galaxy formation and the role of dark matter. It is still too early to assert whether such a relation can be exactly deduced from a model of large-scale structure formation but such a correlation seems ‘natural’ in a hierarchical model of structure formation.

7.2.2 Dark matter in clusters and groups of galaxies

The first indication of the existence of dark matter was revealed by Zwicky during his study of the Coma cluster in 1930 [60]. Galaxy clusters contain from around a hundred to several thousands of galaxies.

7.2.2.1 Virial theorem

Galaxy clusters are gravitationally bound systems decoupled from the cosmic expansion. One can thus assume that they are in a stationary state. The velocity of the galaxies is small and we can model the cluster as an isolated gas composed of massive particles that interact only through gravitation. The acceleration of a galaxy in the cluster is given by

$$\ddot{x}_i = G_N \sum_{j \neq i} m_j \frac{x_j - x_i}{|x_j - x_i|^3}. \quad (7.158)$$

the gravitational energy is

$$2U = -G_N \sum_{j \neq i} \frac{m_j m_i}{|\mathbf{x}_j - \mathbf{x}_i|}, \quad (7.159)$$

while the total kinetic energy is given by

$$2K = \sum_i m_i \dot{\mathbf{x}}_i^2. \quad (7.160)$$

The second derivative of the total moment of inertia, $I \equiv \sum m_i \mathbf{x}_i^2$, is related to the kinetic energy by

$$\ddot{I} = 4K + 2 \sum_i m_i \mathbf{x}_i \cdot \ddot{\mathbf{x}}_i. \quad (7.161)$$

The equation of motion (7.158) can then be used to deduce that

$$\sum_i m_i \mathbf{x}_i \cdot \ddot{\mathbf{x}}_i = G_N \sum_{j \neq i} m_i m_j \frac{\mathbf{x}_i \cdot (\mathbf{x}_j - \mathbf{x}_i)}{|\mathbf{x}_j - \mathbf{x}_i|^3} = U,$$

the second equality being obtained after symmetrization. We thus obtain the relation known as the virial theorem

$$\boxed{\ddot{I} = 2U + 4K.} \quad (7.162)$$

For a stationary system, the value of the time average of the moment of inertia is constant so that $\langle \ddot{I} \rangle = 0$, in which case one has

$$\langle U \rangle + 2\langle K \rangle = 0. \quad (7.163)$$

One can express the mean kinetic energy as $\langle K \rangle = \frac{1}{2} M \langle v^2 \rangle$ and the mean potential energy as $\langle U \rangle = \alpha G_N M^2 / r_h$, where M is the total mass, α is a numerical coefficient that depends on the geometry and on the density profile ($\alpha \sim 0.4$ for galaxy clusters) and r_h the radius in which half of the mass is contained. The virial theorem can thus allow us to estimate the mass of the structure

$$M_{\text{vir}} = \frac{\langle v^2 \rangle r_h}{\alpha G_N}. \quad (7.164)$$

7.2.2.2 Coma cluster

The Coma cluster has a mean redshift of $z = 0.0232$, which places it at around 100 Mpc of the Galaxy. The velocity dispersion along the line of sight is $880 \text{ km} \cdot \text{s}^{-1}$, that is a mean square velocity of $\langle v^2 \rangle \simeq 2.32 \times 10^{12} \text{ m}^2 \cdot \text{s}^{-2}$ assuming that the velocity distribution is isotropic. The estimate of r_h is difficult, mainly because the distribution of dark matter is unknown. Assuming a ratio M/L constant with the radius and a spherical geometry, one can estimate that $r_h \sim 1.5 \text{ Mpc}$, which gives an estimate of the mass of the Coma cluster of $M \sim 2 \times 10^{15} M_\odot$. Since its luminosity is of the order of $L \sim 8 \times 10^{12} M_\odot$, this means that $Q_{\text{Coma}} \sim 250 Q_\odot$. A refined analysis actually gives

$$Q_{\text{Coma}} \sim 400 h Q_\odot, \quad (7.165)$$

and a similar study in the Perseus cluster gives $Q_{\text{Perseus}} \sim 600 h Q_\odot$.

7.2.2.3 Mass of clusters

There are two other methods to determine the mass of a cluster. The first one, presented in Section 7.1.2.6, uses either strong lensing at the centre ($r \lesssim 200$ kpc), which gives constraints of the order of

$$Q_{\text{Coma}} \sim (100 - 200) Q_{\odot}, \quad (7.166)$$

or measurements of the gravitational distortions of the background stars ($r \lesssim 1$ Mpc), which leads to

$$Q_{\text{Coma}} \sim 300 Q_{\odot}. \quad (7.167)$$

Another method relies on the X-ray emission of the hot radiation from the intracluster gas. The presence of such a hot gas captured at the centre of the cluster is proof of the existence of dark matter creating a gravitational well strong enough to trap the gas. Without dark matter, the hot gas would be diluted beyond the cluster in a time smaller than the Hubble time. To estimate the dark-matter mass, let us assume that this gas is in hydrostatic equilibrium and that its pressure compensates gravity

$$\frac{dP}{dr} = -\frac{G_N M(< r) \rho_{\text{gas}}(r)}{r^2}. \quad (7.168)$$

The pressure of the gas is obtained by the measurement of its temperature

$$P = \rho_{\text{gas}} \frac{k_B T}{\mu m_p}, \quad (7.169)$$

where μ is the effective mass of its constituents in units of the mass of the proton. As long as μ is constant, i.e. if the chemical composition is uniform, we have

$$\frac{G_N M_X(< r)}{r} = -\frac{kT(r)}{\mu m_p} \left(\frac{d \ln \rho_{\text{gas}}}{d \ln r} + \frac{d \ln T}{d \ln r} \right). \quad (7.170)$$

The temperature profile is often difficult to measure and the cluster is often assumed to be isothermal and composed of a gas with a radial distribution following a so-called β -profile, $\rho_{\text{gas}} = \rho_0 [1 + (r/r_c)^2]^{-3\beta/2}$. These hypotheses are usually satisfied for clusters with quasi-spherical geometry.

The analysis of the Coma cluster with this method gives a mass of the order of $M_X \sim (3 - 4) \times 10^{14} M_{\odot}$ in a region of 0.7 Mpc around the centre and of $\sim (1 - 2) \times 10^{15} M_{\odot}$ in a region of 3.6 Mpc. These measurements are compatible with the mass determined from the virial theorem and by gravitational lensing. A similar analysis with the Perseus cluster gives $Q = (213 \pm 60) h Q_{\odot}$.

7.2.2.4 Summary

The above-mentioned studies [61] allow us to show that, for clusters,

$$\frac{M_{\text{gas}}}{M_{\text{DM}}} = (15 \pm 5)\%, \quad \frac{M_{*}}{M_{\text{DM}}} = (3 \pm 2)\% \quad (7.171)$$

and the density profiles follow NFW profiles.

To summarize, for $r \gtrsim 1$ Mpc, we do not know the matter distribution. For $10 \text{ kpc} \lesssim r \lesssim 100 \text{ kpc}$, the dynamical mass, the mass deduced from gravitational lensing and that inferred from X-ray emission mass differ approximately by a factor of 2, probably related to projection effects. For $100 \text{ kpc} \lesssim r \lesssim 1 \text{ Mpc}$, the three kinds of observations give the same value for the mass and the matter distribution is compatible both with a NFW profile and an isothermal sphere with core radius. For $r \lesssim 10 \text{ kpc}$, there are insufficient observational diagnostics.

7.2.3 Cosmological evidence

The evolution of homogeneous space-time and the analysis of the evolution of the perturbations in the linear regime provide two additional arguments for the existence of dark matter (see Fig. 7.19).

- Primordial nucleosynthesis gives access to the amount of baryonic matter in our Universe (see Chapter 4).
- The shape of the power spectrum of the matter fluctuations depends on the ratio between the baryonic and dark densities, Ω_{b0}/Ω_{c0} . In particular when the relative quantity of baryons increases, oscillations appear in the matter power spectrum. These oscillations are related to the coupling between baryonic matter and radiation (see Chapter 5).
- The joint analysis of the results of different cosmological observables, for instance, the cosmic microwave background, type Ia supernovæ and gravitational lensing, indicates that around 30% of the matter must be dark.

These analyses allow us to establish the two constraints

$$\Omega_{b0}h^2 = 0.0224 \pm 0.0009, \quad \Omega_{c0}h^2 = 0.113 \pm 0.009. \quad (7.172)$$

These measurements imply that

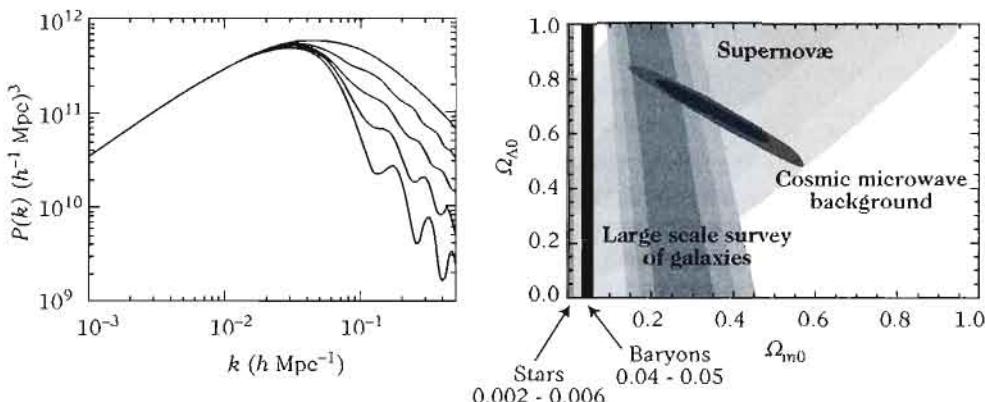


Fig. 7.19 (left): The matter power spectrum of large-scale structures is sensitive to the amount of baryons, ρ_b/ρ_c increasing from the top to bottom curve. (right): The combined analysis of cosmological observations also indicates the necessity for 30% of dark matter.

$$\frac{\rho_c}{\rho_b} = 4.83 \pm 0.87 \quad (7.173)$$

on cosmological scales.

7.2.4 Summary

Table 7.3 summarizes the estimates for the mass/luminosity ratio of various systems and the quantity of dark matter that can be deduced from these systems.

These analyses show that it is difficult to understand the structures of the Universe, regardless of their scale, without introducing dark matter. Much still remains to be understood concerning its physical nature and distribution, in particular its correlation with ordinary matter. As shown in Table 7.4, however, on most scales the quantity of baryonic matter is only a fraction of that of dark matter that, furthermore, seems to decrease with the scale.

Table 7.3 Summary of the estimated mass/luminosity ratio for various systems and implications for the dark-matter content at different scales.

| System | | Scale | Q/Q_\odot | Ω_{c0} |
|---------------------|--------------------|----------|---------------|---------------|
| Milky Way | Oort limit | 100 pc | 5(?) | $0.003h^{-1}$ |
| | rotation curves | 20 kpc | 10 | $0.006h^{-1}$ |
| | satellites | 80 kpc | 40 | $0.024h^{-1}$ |
| | Magellan currents | 100 kpc | 180 | $0.05h^{-1}$ |
| Elliptical galaxies | tuning of the core | 2 kpc | $12h$ | 0.007 |
| | X halo | 100 kpc | 1750 | $0.46h^{-1}$ |
| Spiral galaxies | rotation curves | 50 kpc | 130h | 0.018 |
| Groups | local group | 800 kpc | 100 | $0.06h^{-1}$ |
| | others | 1 Mpc | 260h | 0.16 |
| Clusters | Coma (B band) | 2 Mpc | 400h | 0.25 |
| | Perseus | 2 Mpc | 600h | 0.37 |
| Cosmology | | 3000 Mpc | $0.113h^{-2}$ | |

Table 7.4 Relative ratio of dark and baryonic matter. These constraints indicate that the ratio varies little as a function of the scale of the considered structure.

| System | Spiral galaxy | Cluster | Observable Universe |
|--------------|-----------------------------------|-----------------------------------|---------------------------------|
| DM vs baryon | $\Sigma_c/\Sigma_b = 8.5 \pm 1.5$ | $\Sigma_c/\Sigma_b = 7.2 \pm 2.0$ | $\rho_c/\rho_b = 4.83 \pm 0.87$ |

7.2.5 Candidates and constraints

Dark matter is said to be *hot* or *cold* depending on whether its constituents are relativistic or not at the time of decoupling from the cosmic plasma (see Chapter 4). The only example of dark matter whose existence has been proven is the neutrino, which is a hot dark-matter component. We know, however, that neutrinos cannot account for the whole dark-matter component because they are too light.

We now address the two following aspects: (1) from a structure-formation point of view, what must be the characteristics of this dark matter and (2) from a particle-physics point-of-view, which are the candidates.

7.2.5.1 Cosmological phenomenology

Standard cold dark-matter model

The standard cold dark-matter (SCDM) model assumes that dark matter is composed of massive particles that interact very weakly with each other and with the other components of the Universe. These particles are generically called WIMPs (weakly interacting massive particles).

If these particles are thermal relics, their abundance depends on two parameters, their mass and their cross-section. In general, these two parameters determine the nature of dark matter (cold or hot). For cold dark matter, we have established [see (4.76)] that the relic density is given by

$$\Omega_c h^2 \sim 0.3 \left(\frac{\langle \sigma v \rangle}{10^{-26} \text{ cm}^3 \cdot \text{s}^{-1}} \right)^{-1}. \quad (7.174)$$

From a cosmological point of view, this model reduces to the addition of a pressureless (i.e. collisionless) fluid, thus with the equation of state $w = 0$, coupled to ordinary matter only through gravity. One of the motivations to consider such a model lies in the growth of the cosmological perturbations (see Chapter 5). As we have seen, in the matter-dominated era, sub-Hubble fluctuations grow as a . It is then difficult to understand how the density contrast can become of order unity today if it were of order 10^{-5} at the time of recombination at $z \sim 10^3$. Dark matter resolves this problem since, as it is not coupled to radiation, it can start to develop potential wells before the matter-radiation equality.

This dark-matter model has been extended to the so-called Λ CDM model, where dark matter only represents 30% of the total amount of matter. Today it offers a good understanding of the linear regime and is in agreement with most of the cosmological observations (cosmic microwave background, large-scale structures, etc.). It is at the basis of the *concordance model of cosmology*.

This dark-matter model is considered successful for three main reasons. (1) Its predictions are in agreement with most of the observations of large-scale structures. (2) There are many candidates motivated by high-energy physics. (3) If dark-matter particles interact via the weak interaction, then their relic density has the correct order of magnitude [cf. (7.174)].

The SCDM model is a *hierarchical* model of structure formation (see Section 5.5.2) in which the small structures are formed before the large ones. So the distribution of small halos should not be strongly correlated with that of the larger structures that form later. The properties of the dark-matter halos in this model have been intensively studied by means of N -body numerical simulations. They have made it possible to establish that the dark-matter halos density profiles take the form

$$\rho(r) = \frac{\rho_s}{(r/r_s)^\gamma [1 + (r/r_s)^\alpha]^{(\beta-\gamma)/\alpha}}. \quad (7.175)$$

These profiles behave as $r^{-\beta} = r^{-3}$ at large scales, unlike the behaviour as r^{-2} of the universal profiles. The empirical behaviour as r^{-2} , obtained from the rotation curves, is only an approximation in the *transitionary* region between the behaviour as $r^{-\gamma}$ at the centre and $r^{-\beta}$ at the exterior. Independently from other cosmological details, this model predicts the formation of triaxial halos with very dense cores and

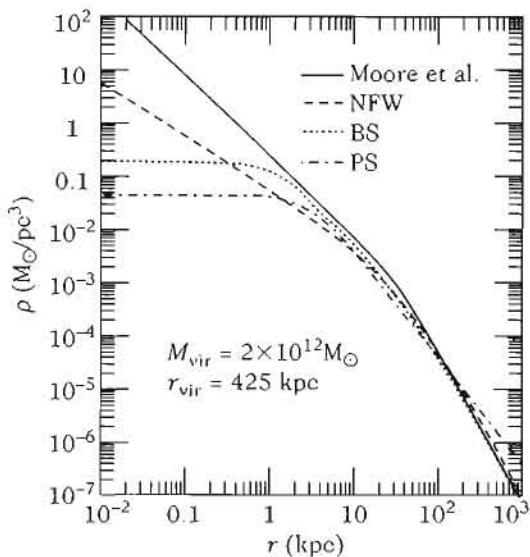


Fig. 7.20 Comparison between the halo density profiles obtained by numerical simulations and the universal density profiles obtained from a fit to observations.

many substructures. These substructures are natural in any hierarchical model of structure formation. The parameters of these profiles ($\alpha, \beta, \gamma, r_s$) are given in Table 7.5. The characteristic density ρ_s is obtained from the total mass and the size of the galaxy. Figure 7.20 compares the different profiles for a typical galaxy.

Table 7.5 Parameters of the dark-matter density profiles (7.175) obtained by different studies of N -body numerical simulations.

| Model | α | β | γ | r_s (kpc) |
|-----------------------------|----------|---------|----------|-------------|
| Navarro-Frenk-White (NFW) | 1 | 3 | 1 | 20 |
| Moore | 1.5 | 3 | 1.5 | 28 |
| Kravstov | 2 | 3 | 0.4 | 10 |
| Isothermal with core radius | 2 | 2 | 0 | 3.5 |

The most recent simulations [59] give other profiles with a form close to the deprojected Sersic profiles

$$\rho(r) = \rho_0(a) \exp\left[-2\mu\left(\frac{r}{a}\right)^{1/\mu}\right] \exp(2\mu), \quad (7.176)$$

with $\mu = 6 \pm 0.2$ and a being the radius where the slope of the profile is equal to -2 . Moreover, in these models the profile $\rho_0(a)$ behaves as $\rho_0(a) \propto 1/\sqrt{a}$.

The value, γ , of the slope of the profile in the inner part of the halo is the source of much debate. These simulations seem to show that the profile does not behave as a power law at the centre. Notice also that these profiles represent a fit to the results of several numerical simulations and do not quantify the dispersion around this fit.

The CDM model must, however, face unresolved problems at small scales.

- It predicts too many substructures, small halos and galaxies in orbit around larger structures, than what is observed.
 - The number of these halos varies as the inverse of the halo mass. Numerous satellites, such as the Magellan cloud for our Galaxy, should thus be observed. The observation of the Local Group reveals less than a hundred galaxies (36 dwarf galaxies are known in a radius of 0.5 Mpc) while numerical simulations and analytical estimates predict around a thousand satellites.
 - These satellites should induce a more significant thickening of the galactic disk than is observed.
- The predicted density profiles of the dark-matter halos are more peaked than what is observed in many self-gravitating systems.
 - Clusters, observed via gravitational lensing, have less-peaked cores than that predicted for massive halos. Notice that the observations of the core are, however, difficult since the baryons dominate. Let us also stress that the interpretation of the rotation curves assumes an axial symmetry, while in these regions the stars are not always on Keplerian orbits.
 - The central density of the dwarfs and low surface brightness galaxies is too weak and the density profile too smooth compared to that obtained from numerical simulations. Surprisingly, the smallest galaxies (e.g. DD0210, DD0154, etc.) are consistent with a quasi-constant dark-matter profile.
 - The persistence of the bars in galaxies of high surface brightness also implies that these galaxies have cores of small density.
 - If the core becomes dense too early, then the cooling of the baryons could be too efficient, which would lead to galactic disks with angular momenta one order of magnitude too small.
 - Dwarf spheroidal galaxies are also dominated by dark matter, but they show no sign of a peaked profile.
- Dwarf galaxies of weak luminosity seem to be more strongly correlated with luminous galaxies than they should. In particular, they do not fill the big voids between the distribution of luminous galaxies and localize on the borders of these voids.
- The observation of galaxy formation at $z \sim 3 - 4$, later than that of the most massive bright galaxies, seems in contradiction with a hierarchical model of structure formation. Similarly, there are supermassive black holes ($10^9 - 10^{10} M_\odot$) at high redshifts, which does not seem very understandable in this scenario.

Nevertheless, one should not draw too hasty conclusions. One of the keys to the debate lies in the behaviour at large distances where simulations predict that the profiles behave as $1/r^3$. These external zones are, to date, difficult to study observationally. Two sources of information can be considered: First, galaxy-galaxy lensing is sensitive to the most massive objects and it seems [62] that observations of the cosmic shear as a function of the distance for galaxies of mass larger than $10^9 M_\odot$ is compatible with the NFW profiles. Second, data from the SDSS survey have allowed for the study of satellite galaxies of various clusters [63], the behaviour of which seems compatible with a NFW profile. Both these observations do not establish the behaviour of the density profile at large distances in a definitive way but indicate a tendency to be confirmed.

Other models

The problems of the SCDM model at small scales should not overshadow its successes at large scales and in the linear regime. These difficulties have motivated the development of many phenomenological models of dark matter with the main objective of modifying the properties of the structures at small scales. Each one of these models brings new responses to one or several problems of the SCDM model and has many proper signatures that future observations and numerical simulations should be able to discriminate. The following list is not at all exhaustive but illustrates the variety of attempts.

1. *Hot dark matter* (HDM): this model was motivated by the study of neutrinos. This form of dark matter is relativistic at the time of decoupling and at the time of the large-scale structure formation. Its main problem is that dark matter particles are too fast to be trapped and to collapse gravitationally. Also, the formation of structures starts too late since equality is delayed. An important scale is the free streaming scale λ_{fs} of a relativistic particle before it becomes non-relativistic. In HDM models, fluctuations on scales smaller than λ_{fs} are erased so that the spectrum of matter perturbations has a well-defined cutoff at λ_{fs} (see Fig. 5.16). For instance, for massive neutrinos, $\lambda_{\text{fs}} \sim 40(30 \text{ eV}/m_\nu) \text{ Mpc}$. This implies that protogalaxies with mass of the order of $10^{15} M_\odot$ are the first objects to form. Smaller objects then form by *fragmentation*. The model of structure formation is then very different from the hierarchical one of the SCDM model.
2. *Warm dark matter* (WDM): this model [64] is a compromise between the HDM and CDM models. Dark matter would be composed of relativistic thermal relics at the time of decoupling but that would become non-relativistic before equality. The characteristic distance below which density fluctuations are erased is given by

$$R \sim 0.2 (\Omega_{\text{co}} h^2)^{1/3} \left(\frac{m}{1 \text{ keV}} \right)^{-4/3} \text{ Mpc}. \quad (7.177)$$

A warm dark-matter candidate must therefore have a mass of the order of 0.1–1 keV.

The main predictions of this model are (1) the smoothing of the core of the halo, which translates into a decrease in the core density and an increase in the core radius, (2) a decrease in the characteristic density of small mass halos, (3) a

global reduction of the number of small mass halos, (4) a decrease in the number of small mass halos inside larger mass halos, (5) the formation of voids containing no small mass halos, (6) the formation of small halos at redshift $z < 4$ via a fragmentation process and (7) the suppression of the halo formation at large redshifts ($z > 5$). The properties (1) and (5) provide solutions to some of the SCDM problems while (6) and (7) offer the possibility of specific tests. If the delay in the structure formation was too large, the model would then become incompatible with observations.

3. *Strongly self-interacting dark matter (SIDM)*: this model [65] supposes that dark matter has a self-interaction cross-section that cannot be neglected while the annihilation rate remains very weak. This interaction only alters the dynamics of dark matter when its density becomes important, so that the cosmic microwave background and the large-scale structures remain insensitive to this modification. Concretely, the mean free path of SIDM should range between 1 kpc and 1 Mpc when the density is of the order of $0.4 \text{ GeV} \cdot \text{cm}^{-3}$ (typical density at the position of the Solar System). Indeed, if the mean free path was larger, no interaction would occur on scales of the size of the halo and if it was smaller than 1 kpc, dark matter would behave as a collisional gas, which would greatly modify the evolution and the structure of the halos. The mass of these particles must range between 1 MeV and 10 GeV and its cross-section be of the order of

$$\sigma \sim 8 \times 10^{-25} \left(\frac{m}{1 \text{ GeV}} \right) \left(\frac{\lambda}{1 \text{ Mpc}} \right)^{-1} \text{ cm}^2. \quad (7.178)$$

The halo structure is only modified above a given surface density and induces a matter redistribution in these regions, thus producing a less dense and more spherical core. Notice that if σ is too large then the halo can completely evaporate, which can lead to catastrophic effects, but would be observable without any difficulty in the halo demography.

4. *Repulsive dark matter (RDM)*: this model [66,67] supposes that the dark matter is a massive boson condensate with a short-distance repulsive potential. The internal part of the halo then behaves as a superfluid, which has the tendency to smooth out its profile. The halo cores would then have a minimum size independent of their density. In the linear regime, scales smaller than a few Mpc are suppressed.
5. *Fuzzy dark matter (FDM)*: this model [68] supposes that dark matter takes the form of a very light scalar particle with a Compton wavelength of the order of the galaxy core. At small scales, the wave-like behaviour of dark matter prevents gravitational collapse so that the core of the halo is smoothed. The Jeans length, under which density perturbations are stable (see Chapter 5), is given by

$$\lambda_J \sim 55 \left(\frac{m}{10^{-22} \text{ eV}} \right)^{-1/2} \left(\frac{\rho}{2.8 \times 10^{11} M_\odot \text{ Mpc}^{-3}} \right)^{-1/4} \text{ kpc}. \quad (7.179)$$

A field of mass $m \sim 10^{-22}$ eV can then provide a smooth density profile on scales of the order of the kpc. This also induces a sudden cutoff around the

wavenumber $k \sim 4.5(m/10^{-22} \text{ eV})^{1/2} \text{ Mpc}^{-1}$ in the linear power spectrum, which can affect the abundance of small-mass objects. This mass scale can, however, seem artificial in the framework of particle physics. Notice that the mass of a particle with a Compton wavelength of the size of the observable Universe is of the order of $m \sim 10^{-33} \text{ eV}$ and that a mass of this order can also be related to the acceleration of the Universe (cf. Chapter 12).

6. *Self-annihilating dark matter* (SADM): this model [69] supposes that dark matter can annihilate itself in the strong density regions where two-body interactions become important. The annihilation produces a radiation that should not include photons, as their flux would exceed the observational constraints. The annihilation has the effect of decreasing the core density and to induce a global expansion of the halo due to the decrease in the gravitational potential. The annihilation rate depends on the density and on the velocity dispersion, $\Gamma = n\langle\sigma v\rangle$. In the particular case where $\Gamma = (\rho/m)\sigma$, the density profile takes the form

$$\rho(r) = \rho_s \left[\frac{r}{r_s} \left(1 + \frac{r}{r_s} \right)^2 + \frac{\rho_s}{m/(\sigma t_0)} \right]^{-1}, \quad (7.180)$$

where t_0 is the age of the Universe. A cross-section $\langle\sigma v\rangle \sim 10^{-29} (m/1 \text{ GeV}) \text{ cm}^2$ gives a core density of 1 GeV/cm^3 , in agreement with observations. The effects on the number of small halos has not been studied.

7. *Decaying dark matter* (DDM): if the dense halos can decay into relativistic particles and into particles of smaller mass, then the density of cores formed earlier can be noticeably reduced without affecting the large-scale structure [70]. A potential problem of this model is that the production rate of dense and massive clusters must be far more important than in the SCDM model to obtain a correct distribution after the decay of dark matter.
8. *Massive black holes* (BH): if the main part of dark matter was composed of massive black holes of around 10^3 solar masses, the dynamical friction between black holes and ordinary matter could lead to the migration of these black holes towards the centre. They can thus give rise to the central black hole of the galaxy, decreasing the quantity of dark matter in the core. But we have to assume that these black holes are present to begin with.
9. *Initial spectrum*: the breaking of the scale invariance of the primordial fluctuations power spectrum at around $k = 5 - 10 h^{-1} \text{ Mpc}$ can marginally modify the structures of halos at small scales [71].

These models introduce two kinds of modifications compared to the SCDM model.

- (1) The halo is only affected when the interaction rate becomes larger than a critical value (SIDM, BH, SADM). (2) The models have a characteristic time or distance scale below which the halo is modified (WDM, MDM, FDM, DDM). They therefore lead to different observational predictions. For instance, to yield the same structures today, the objects of a given mass must form earlier in DDM, SADM and BH while objects

of smaller mass form later in FDM and HDM and by fragmentation in WDM. The demography of the halos and satellites is modified together with the core density profile. To illustrate these examples, Fig. 7.21 summarizes quantitatively the implications of these scenarios. The central structure of the halos, and mainly that of the dwarfs and galaxies of low surface brightness, seem to be the key to characterize dark matter.

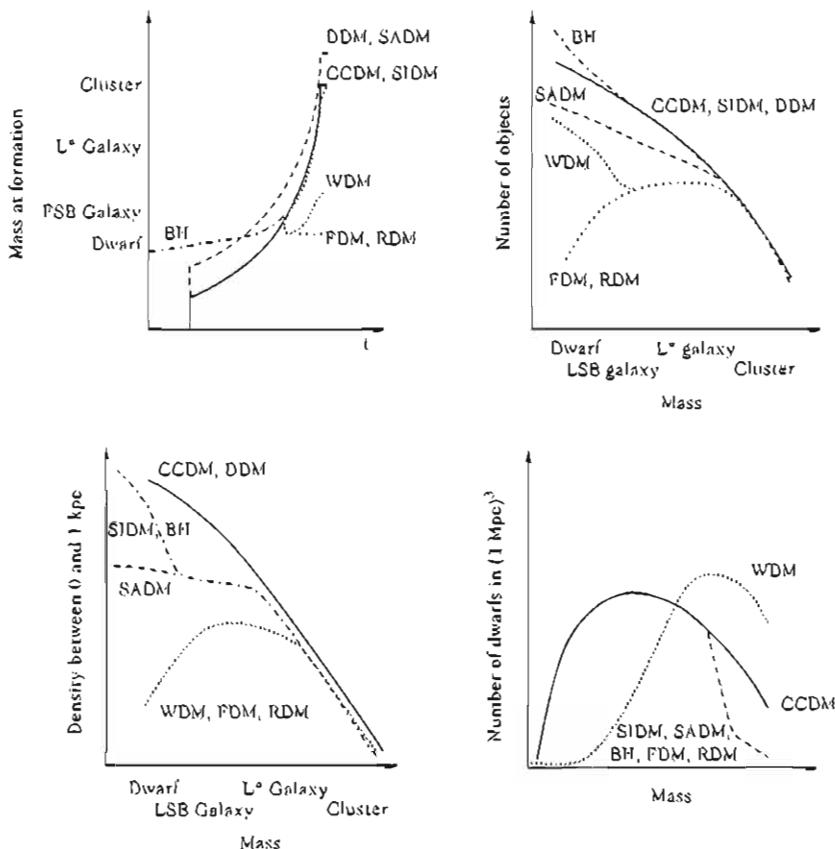


Fig. 7.21 Predictions and signatures for various models of dark matter. (top left): the formation time of a structure with a given mass for structures of increasing mass (dwarfs, low surface brightness, galaxies, clusters). (top right): dependence of the number of objects as a function of its mass. (bottom left): density in the inner 1 kpc region as a function of the mass of the system. (bottom right): number of dwarfs in a volume of $(1 \text{ Mpc})^3$ as a function of the mean density in this volume. From Ref. [72].

7.2.5.2 High energy candidates

The standard model of particle physics (see Chapter 2) does not provide any dark-matter candidate. The only (stable) neutral fermion, the neutrino, has a vanishing

mass, although it is known experimentally to be massive. Nevertheless, almost all the extensions of this standard model provide several candidates, and in particular massive neutrinos.

For detailed reviews concerning these candidates, detection methods and experimental constraints, see Refs. [73–76].

We only consider here non-baryonic candidates. As explained in Section 7.1.2.1, microlensing allows us to put a strong constraint on the abundance of brown dwarfs. Moreover, nucleosynthesis and the large-scale structure formation provide us with strong arguments against the fact that all dark matter could be baryonic. dark-matter candidates can be classified according to various criteria.

- The first classification is based on the candidate mass at the time of galaxy formation, at a temperature of the order of $T \sim 1$ keV. Particles of mass $m \ll 1$ keV are called *hot* (such as neutrinos), those of mass $m \gg 1$ keV *cold* (neutralinos, axions,...) and those of mass $m \sim 1$ keV *warm* (sterile neutrinos, gravitinos).
- A second classification can be performed on the basis of the production mechanism. One can distinguish between *thermal relics*, which were in thermal equilibrium during the primordial Universe (neutrinos, neutralinos,...) and *non-thermal relics*, produced by some non-thermal mechanisms (axions produced by cosmic strings, WIMPZILLAs,...).
- The last classification relies on the framework of particle physics in which they appear. We can distinguish between *existing* candidates (neutrinos), candidates that are *motivated* but not yet observed (e.g. light supersymmetric particles, axion), i.e. that naturally appear in the framework of particle physics independently from the dark-matter problem and that have masses and cross-sections leading to relic densities in agreement with the constraints, and *exotic* candidates that can, nonetheless, be interesting ideas (e.g., WIMPZILLAs). This last classification is indeed more arbitrary and variable as a function of the progress in particle physics.

7.2.5.3 Neutrinos

Many, if not all, extensions to the standard model of particle physics allow for the inclusion of a massive neutrino with mass $m_\nu < 100$ eV.

Neutrinos oscillations experiments (see Ref. [74]) allow us to establish that

$$\nu_\mu \leftrightarrow \nu_\tau : \quad \Delta m_{23}^2 \sim 3 \times 10^{-3} \text{ eV}^2, \quad \nu_e \leftrightarrow (\nu_\mu, \nu_\tau) : \quad \Delta m_{12}^2 \sim 7 \times 10^{-5} \text{ eV}^2. \quad (7.181)$$

This leads to the conclusion that the mass of the heaviest neutrino must be larger than 0.05 eV and laboratory experiments show that the mass of each of the three neutrinos must be lower than 2.8 eV.

Neutrinos thus contribute to hot dark matter and their abundance (see Chapter 4) can be evaluated as

$$\Omega_{\nu 0} h^2 = \sum g_i \left(\frac{m_i}{90 \text{ eV}} \right), \quad (7.182)$$

with $g_i = 1$ for Majorana neutrinos and $g_i = 2$ for Dirac neutrinos. So with only one Majorana neutrino, the constraint on the neutrino mass implies that

$$\Omega_{\nu 0} h^2 < 0.0006. \quad (7.183)$$

The analysis of data from the cosmic microwave background, from primordial nucleosynthesis and large structures implies that

$$\Omega_{\nu 0} h^2 < 0.0076, \quad (7.184)$$

which requires that $\sum g_i m_i < 0.7 \text{ eV}$. To conclude, the combination of laboratory and cosmological constraints gives

$$0.0006 < \Omega_{\nu 0} h^2 < 0.0076, \quad 0.05 \text{ eV} < \sum_i m_i < 0.7 \text{ eV}. \quad (7.185)$$

These constraints imply that neutrinos cannot represent the main part of dark matter (for a review on cosmological neutrinos, see [77, 78]). We are therefore led to consider particles whose existence is only suggested from the extensions to the standard model of particle physics.

7.2.5.4 Other kinds of neutrinos

Neutrinos are fermions. The Pauli exclusion principle implies that light neutrinos cannot represent the main part of dark matter. Indeed, in a dwarf galaxy, which must be dominated by dark matter, these neutrinos must be confined in phase space, with a volume determined by the Pauli principle. One can show that in order to explain the distribution of dark matter in dwarf galaxies, we need that the mass m_ν of at least one species of neutrino satisfies [79]

$$m_\nu \geq 120 \left(\frac{\sigma_v}{100 \text{ km} \cdot \text{s}^{-1}} \right)^{-1/4} \left(\frac{r_c}{1 \text{ kpc}} \right)^{-1/4} \text{ eV}, \quad (7.186)$$

which is in contradiction with the constraints (7.185).

Sterile neutrinos [80], that are right-handed neutrinos coupled only to left-handed neutrinos (active neutrinos), have also been considered. These neutrinos are mainly produced by neutrino oscillations. Depending on their masses, they can behave as hot, warm or cold dark matter. As an example, the mass term for a model with only one generation of right-handed neutrinos is given by the Lagrangian

$$\mathcal{L} = \mu \left(\frac{\phi}{v} \right) \bar{\nu}_L \nu_R + M \nu_R \nu_R + \text{h.c.}, \quad (7.187)$$

where ϕ is the Higgs field and v its vacuum expectation value (VEV). The case of hot dark matter corresponds to $\mu \sim 90 h^2 \text{ eV}$ and $M \ll \mu$ or $\mu^2/M \sim 90 h^2 \text{ eV}$ and $M \gg \mu$. When sterile neutrinos are candidates for dark matter, M is the relevant mass.

7.2.5.5 Axions

The combined action of charge conjugation and parity (see Chapter 2) is not an exact symmetry even in the standard model of particle physics described in Chapter 2. This

$\hat{C}\hat{P}$ violation in the strong interaction sector is expected in the framework of quantum chromodynamics. The effective Lagrangian then contains a term of the form [81]

$$\mathcal{L}_\theta \propto \theta g_3^2 G_{\mu\nu}^a \epsilon^{\mu\nu\rho\sigma} G_{a\rho\sigma}, \quad (7.188)$$

$G_{\mu\nu}^a$ being the tensor of the gluons. This indeed violates parity (\hat{P}) and time reversal (\hat{T}) and therefore $\hat{C}\hat{P}$ as $\hat{C}\hat{P}\hat{T}$ must be conserved. Such a term induces a permanent electric dipole for the nucleons, which is, however, not observed down to the current experimental sensitivity, corresponding to the constraint $\theta < 10^{-10}$.

Axions were introduced by Peccei and Quinn [82] to solve this problem. For this, the symmetry group of the standard model is extended by a global symmetry, $U(1)_{PQ}$, which is spontaneously broken. The symmetry being global, its Goldstone boson cannot be absorbed by the Higgs degree of freedom, and so there must exist an additional scalar field, called the axion, a , which is associated with the phase of the $U(1)_{PQ}$ transformation. Notice that such additional global symmetries $U(1)$ also appear in some compactification schemes of type I/II/heterotic string theories.

The phenomenology of the axions is determined by a unique parameter, f_a which characterizes the scale of the Peccei–Quinn symmetry breaking, and in terms of which the axion mass is given by

$$m_a = 0.62 \left(\frac{10^7 \text{ GeV}}{f_a} \right) \text{ eV}. \quad (7.189)$$

Moreover, one can show that the axion is coupled to two photons since its effective action contains a term of the form

$$\mathcal{L}_{a\gamma} = \frac{\alpha}{f_a} F_{\mu\nu} \tilde{F}^{\mu\nu} a \quad (7.190)$$

where $F_{\mu\nu}$ is the Faraday tensor and $\tilde{F}_{\mu\nu} = \frac{1}{2} \epsilon_{\mu\nu\rho\sigma} F^{\rho\sigma}$, its dual. This coupling implies that the axion can decay into two photons. This process dominates unless $m_a > 2m_e$ as the axion can then rapidly decay into electron–positron pairs.

Moreover, in the presence of a magnetic field, an axion can oscillate into a photon of energy m_a . Laboratory and astrophysical constraints imply that $m_a < 0.01$ eV. Its coupling to other fields is thus very weak and the axion therefore behaves as cold dark matter. The original Peccei–Quinn axion is thus very constrained, mainly since it interacts with ordinary matter and because one needs $f_a \sim 250$ GeV. The properties of the axion have been extended in order to introduce it into other sectors, making it an *invisible axion*.

7.2.5.6 Candidates motivated by supersymmetry

From a theoretical point of view, the best motivated dark matter candidates are probably those that appear in the framework of supersymmetric theories (see Chapter 10).

In order to be a WIMP candidate, the lightest supersymmetric particle (LSP) must be electrically neutral and a colour singlet.

- *sneutrino*: the supersymmetric partner of the neutrino, $\tilde{\nu}$, has been considered as a good candidate for a long time and its relic density has the correct order

of magnitude if its mass is of the order of $m_\nu = 550 - 2300$ GeV. Nevertheless, the cross-section of its interaction with nucleons exceeds constraints obtained by dark-matter detection experiments that impose that $m_\nu \gtrsim 20$ TeV.

- **gravitino:** the supersymmetric partner of the graviton can be the LSP in locally supersymmetric theories (supergravity). Nevertheless, since it only interacts gravitationally, it is almost impossible to detect. The gravitino has a spin equal to 3/2 and gets a mass, $m_{3/2}$, when supersymmetry is broken. Since its interaction is purely gravitational, its interaction rate is of the order of

$$\Gamma_{3/2} \sim G_N^2 T^5 \sim \frac{T^5}{M_P^4}. \quad (7.191)$$

Decoupling occurs when $\Gamma_{3/2} \sim H \sim T^2/M_P$, i.e. at around $T \sim M_P$. Using the relation (4.80), we expect a relic density of the order of $\Omega_{3/2} h^2 \sim 2.25(m_{3/2}/10\text{ eV})$ [using $g_{3/2} = 4$ and evaluating $q_*(x_f) \sim 200$]. If $h \sim 0.7$, this requires that

$$m_{3/2} \lesssim 4.42 \text{ eV}. \quad (7.192)$$

This value is too small since $m_{3/2}$ controls the supersymmetry-breaking scale and such a value does not allow supersymmetric particles to have a mass larger than a TeV. This is what is called the *gravitino problem*. It can be avoided by assuming that the gravitino is heavier and that it decays during the cosmic evolution. The decay rate is then of the order of $\Gamma_{3/2} \sim m_{3/2}^3/M_P^2$ instead of (7.191). It becomes comparable to the expansion rate, $H \sim T^2/M_P$, at a temperature of the order of $T_d \sim m_{3/2}^{3/2}/M_P^{1/2}$ which corresponds to the temperature at which gravitinos decay. The Universe is then reheated and constraints are then set by nucleosynthesis as this decay should not modify the light-element abundances. Finally, gravitinos could be overproduced during the primordial Universe [83].

- **neutralino:** there are four neutralinos, χ_n^0 with $n = 1 \dots 4$ in order of increasing mass, in the minimally supersymmetric model (MSSM) that are linear combinations of the Wino, \tilde{W}^3 , the Bino, \tilde{B} , and of the two Higgsinos, \tilde{H}_d and \tilde{H}_u , these particles being the supersymmetric partners of, respectively, the gauge bosons W_3 and B , and of neutral Higgs bosons, H_d and H_u . The four neutralinos states are thus given by

$$\chi_n^0 = N_{1n} \tilde{B} + N_{2n} \tilde{W}^3 + N_{3n} \tilde{H}_d + N_{4n} \tilde{H}_u, \quad (7.193)$$

where the coefficients N_{in} are the eigenvectors of the neutralino mass matrix. This mass matrix depends on θ_W and m_Z (see Chapter 2) and four of the 63 undetermined parameters from the MSSM [M_1 and M_2 , the masses of the gauginos of the groups $U(1)_Y$ and $SU(2)_L$, $\tan \beta$, the ratio of the VEVs of the two Higgs fields, and μ , the mass parameter for the Higgsinos].

The problem one has to solve is understanding whether, for a given set of values of these parameters, the cross-sections for the annihilation and interaction with the nucleons (see Ref. [84]), as well as the mass of χ_1^0 , are compatible with the relic abundances. Since neutralinos are non-relativistic, their cross-section can be

expanded as $\langle \sigma v \rangle = a + bv + \dots$. Using the analytical approximation (4.76), we obtain a lower bound for the relic density [84]

$$\Omega_{X0} \gtrsim \left(\frac{m_X}{200 \text{ TeV}} \right)^2. \quad (7.194)$$

This estimate can be modified by a factor of 2 if coannihilations are taken into account. Cosmological and accelerator constraints lead to the conclusion that the neutralino mass must satisfy [73]

$$108 \text{ GeV} \leq m_X \leq 370 \text{ GeV} \quad (\mu > 0) \quad 160 \text{ GeV} \leq m_X \leq 430 \text{ GeV} \quad (\mu < 0). \quad (7.195)$$

- **axino:** the supersymmetric partner of the axion usually behaves as hot or warm dark matter. It can only behave as dark matter if the reheating temperature is small enough.

7.2.5.7 WIMPZILLAs

The hypothesis that dark matter is a thermal relic is somehow very restrictive. It forces the mass of the possible candidate to be lower than a few TeV, i.e. of the order of the electroweak scale. The reason for the existence of such a limit lies in the fact that for a particle of mass m_{DM} , the annihilation cross-section has a maximal bound of the order of m_{DM}^{-2} , called the *unitary bound* [85]. This imposes the mass of the particles [84] to be of the order of

$$m_{\text{DM}} \lesssim 340 \text{ TeV}. \quad (7.196)$$

WMAP measurements have reduced this constraint to

$$m_{\text{DM}} \lesssim 34 \text{ TeV}. \quad (7.197)$$

It can only be avoided if dark matter is not a thermal relic, since then the relations (4.76) and (4.80) cannot be applied.

WIMPZILLAs (X) are ultraheavy relics produced through a non-thermal mechanism. In order for this mechanism to be viable, the particles should be stable, or at least their lifetime larger than the age of the Universe that, for all practical purposes, is equivalent. In particular, the decay of WIMPZILLAs could explain the origin of ultrahigh-energy cosmic rays if their lifetime happened to be comparable with the age of the Universe. Furthermore, we have to require that WIMPZILLAs were not at equilibrium before their abundance froze, since otherwise $\Omega_X \gg 1$. A sufficient condition is that (see Chapter 4),

$$\Gamma_X = n_X \langle \sigma v \rangle < H. \quad (7.198)$$

Reciprocally, if the particles X were produced during the radiation era, at a temperature T_* , then their relic abundance would be $\rho_{X0} = m_X n_X(T_*)[s_0/s(T_*)]$, and therefore that $n_X(T_*)/H_* \sim (\Omega_{X0}/\Omega_{\gamma0}) T_0 M_p T_*/m_X$. Thus, taking into account the

unitary bound $\langle \sigma v \rangle < m_X^{-2}$, we obtain that $\Gamma_X / H_* < (\Omega_{X0}/\Omega_{\gamma0}) T_0 T_* M_p / m_X^3$. The non-equilibrium condition then becomes

$$\left(\frac{m_X}{200 \text{ TeV}}\right)^3 > \left(\frac{T_*}{200 \text{ TeV}}\right). \quad (7.199)$$

To sum up, a non-relativistic particle of mass $m_X \gtrsim 200$ TeV, created at $T_* \lesssim m_X$ with an abundance small enough so that $\Omega_{X0} < 1$, actually has such a small density that it could not have been in equilibrium.

Various scenarios of WIMPZILLAs production have been proposed [86, 87] among which one relies on the amplification of perturbations of a non-minimally coupled scalar field during the reheating phase. Just as the inflaton fluctuations (see Chapter 8), the scalar field $X(x, \eta)$ can be expanded in modes as

$$X(x, \eta) = \int \frac{d^3 k}{(2\pi)^{3/2} a(\eta)} [h_k(\eta) e^{ik \cdot x} \hat{a}_k + \text{h.c.}] . \quad (7.200)$$

The evolution equation of the mode h_k is then

$$h_k'' + \omega^2(k, \eta) h_k = 0, \quad \omega^2(k, \eta) = k^2 + m_X^2 a^2 + (6\xi - 1) \frac{a''}{a}. \quad (7.201)$$

The initial conditions are set in the inflationary era by assuming that

$$h_k(\eta \rightarrow -\infty) \propto \exp \left[-i \int^\eta \omega(k, \eta') d\eta' \right].$$

At the end of inflation, at $\eta = \eta_I$, the field contains modes of positive and negative frequencies,

$$h_k(\eta_I) = \alpha_k \exp \left[-i \int^{\eta_I} \omega(k, \eta') d\eta' \right] + \beta_k \exp \left[+i \int^{\eta_I} \omega(k, \eta') d\eta' \right].$$

Thus, at the end of inflation, particles have been produced. Their density is related to the value of the Bogoliubov coefficient, β_k , by

$$n_X(\eta_I) = a^{-3} \int \frac{d^3 k}{(2\pi)^3} |\beta_k|^2. \quad (7.202)$$

Particles with masses comparable to the Hubble constant at the end of inflation, $m_X \sim H_I$, can be created with a density $\rho_X(\eta_I) = m_X n_X(\eta_I)$. Figure 7.22 describes the time evolution of the Bogoliubov coefficient and the relic density obtained as a function of m_X/H_I . Various scenarios can be constructed depending on the way inflation ends.

7.2.5.8 Other candidates

There is a long list of other candidates [73–76] that we will not detail here. We simply mention the following possibilities.

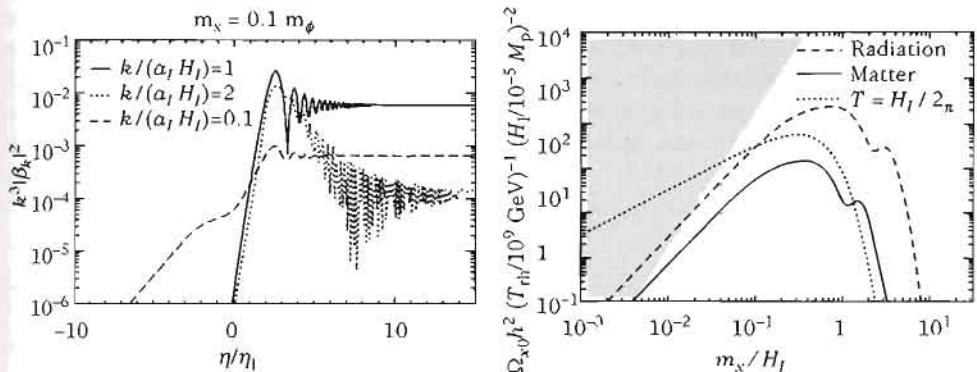


Fig. 7.22 (left): Conformal time evolution of the Bogoliubov coefficient for various wave numbers. η_I corresponds to the end of inflation. (right): Contribution to $\Omega_{\chi 0} h^2$ of gravitationally produced WIMPZILLAs as a function of the ratio between its mass, m_χ , and the Hubble constant at the end of inflation, H_I . T_{rh} is the reheating temperature and the shaded region corresponds to parameters for which thermalization can occur. Inflation is assumed to be smoothly followed either by a matter-dominated era (solid line) or a radiation era (dashed line). The dotted line shows a thermal density with the temperature $T = H_I / 2\pi$. From Ref. [86].

- *Light scalar particle*: the Lee–Weinberg bound [85] can be bypassed for non-fermionic dark matter. A model with a scalar particle having a mass in the range 1–100 MeV has recently been proposed [88]. Although this candidate is relatively *ad hoc* from a particle-physics point of view, it is motivated by the fact that it may offer an explanation of the γ -ray emission at 511 keV observed in the bulge of our Galaxy. This emission would be due to the annihilation of these light scalar particles into positrons, which would then produce the recently observed γ -rays when annihilating.
- *Model of the ‘little Higgs’*: the ‘little Higgs’ model [89] is an alternative model to supersymmetry in which the Higgs from the standard model is a pseudo-Goldstone boson, the mass of which is approximatively protected by a global symmetry (see Chapters 2 and 9). This makes it possible to stabilize the electroweak scale at around 10 TeV. Various scenarios of this model contain scalar particles that could be good dark-matter candidates. In one of the versions [90], a new symmetry is introduced at the TeV scale and implies the existence of WIMPs with a mass of a few TeV.
- *Kaluza–Klein modes*: in theories with higher dimensions compactified à la Kaluza–Klein (see Chapter 13) where all the fields can propagate along the extra dimensions (these models are called *universal extra dimensions*), the Kaluza–Klein modes of the standard model fields can act as dark matter. The lightest Kaluza–Klein particle [91] can be stable if the extra dimension is compactified on an orbifold, such as S^1/Z .

- **Vortons:** vortons are superconducting cosmic string loops stabilised by their current [92]. They are formed during a phase transition associated with spontaneous symmetry breaking. As for monopoles, they have the tendency to dominate very rapidly the matter content of the Universe, which is known as the *vorton excess problem*. Light vortons, however, were also suggested as candidates of ultrahigh-energy cosmic rays.

7.2.5.9 Direct searches

From a phenomenological point of view, WIMPs are characterized by their mass and the cross-section of their annihilation and interaction with the nucleons (see Fig. 7.23). These parameters make it possible to compute the relic density of the candidate so that they are constrained by cosmological observations (see Fig. 7.24 left).

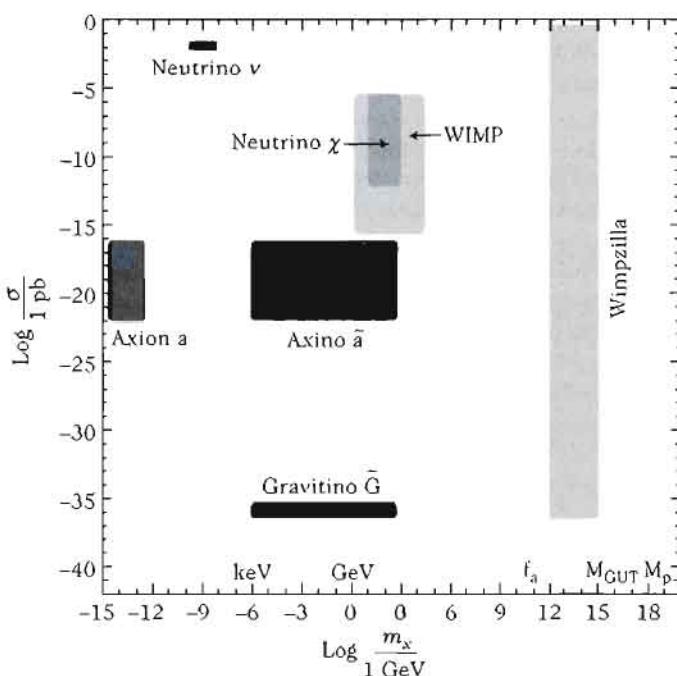


Fig. 7.23 Different dark-matter candidates in parameter space (cross-section, mass). From Ref. [94]

Many experiments [93] have attempted to directly detect the elastic scattering of WIMPs on nucleons. The principle of the experiment is based on the fact that the WIMP scattering induces a recoil of the nucleon that can be detected by three methods. (1) If the detector is a semiconductor (for instance, germanium), the recoil can release free charges through *ionization* (electron–hole pairs). These charges can then be collected by an electric field applied at the boundary of the crystal. (2) Some crystals and liquids release flashes of light when the nucleus slows down. This is the

scintillation phenomena. The quantity of light produced depends on the recoil of the nucleus. The materials used for these detectors are mostly sodium iodide, NaI, and liquid xenon. (3) In crystals, the recoil energy is transformed into vibrations of the crystal lattice (phonons) which can be detected, at very low temperatures, through the increase of temperature that they induce. Bolometers made of sapphire, TeO₂ and germanium, LiF have, for instance, been used. One of the main difficulties of these methods lies in the existence of a natural radiation that is more intense than the source of WIMPs we are looking for (cosmic rays or natural radioactivity), which generates an intensive flux of parasitic events. The incident particles are then mainly photons, electrons and neutrons. To prevent these noises, the experiments are installed in deep underground sites.

Two parameters determine the efficiency of such a method, the incident WIMPs flux and the WIMP-nucleon cross-section. In a given model, this cross-section can be computed and the incident flux is constrained by astrophysical data. In the most optimistic models, the expected rate of events is of the order of 10^{-7} per day per kilogram. We can thus put constraints on these two parameters (see Fig. 7.24). Moreover, when the model is well specified (for instance, in the minimally supersymmetric model) one can put constraints directly on the parameters of this model (in this case on μ , β , M_1 and M_2). These constraints are then compared with accelerator experiments limits.

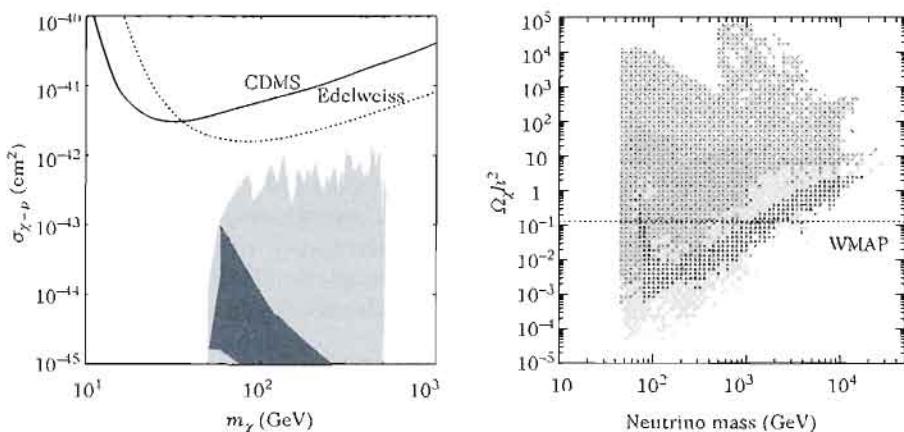


Fig. 7.24 (left): Experimental limits obtained by the CDMS (black) and Edelweiss (grey) experiments in the (cross-section, mass)-plane. The shaded region corresponds to the predictions from the constrained MSSM and the light region to that from the so-called ‘non-universal Higgs mass’ model. From [95]. (right): Predictions of the LSP relic abundance as a function of the mass and the cosmological constraints. For each mass many relic densities are possible as a function of the value of the other parameters of the model. From Ref. [75].

7.2.5.10 Gravitational alternatives

All of the conclusions we have drawn on the existence of dark matter rely on the validity of general relativity to describe gravity, and in fact mostly on the validity of Newton's law. The behaviour of the galaxy rotation curves can, however, be explained without requiring any dark matter if the gravitational potential behaves as

$$\Phi = \frac{G_N M}{r_0} \ln \left(\frac{r}{r_0} \right), \quad (7.203)$$

at large distances, assuming r_0 to be a characteristic scale of the order of a few kpc. In this regime, the velocity of a test body in Keplerian orbit tends to a constant, $v^2 = G_N M/r_0$. So, if the modification of gravity occurs at a fixed distance scale, we find that $L \propto M \propto v^2$, which is in contradiction with the Tully–Fischer relation. This conclusion is generic to any modification associated with a characteristic distance scale. With the definition (7.133), we can rewrite the gravitational potential as

$$\Phi = -\frac{G_N M}{r} + \sqrt{G_N M a_0} \ln \left(\frac{r}{r_0} \right), \quad (7.204)$$

where a_0 is a constant with dimensions of acceleration. We see that the transition between the Newtonian and modified Newtonian regimes has to occur at a scale of order $r \sim \sqrt{G_N M/a_0}$, which depends on the mass of the galaxy.

7.2.5.11 MOND: formulation

As illustrated in Fig. 7.25, there is no correlation between the ratio M/L and the size of the galaxy. The Newtonian ratio M/L is obtained by computing the dynamical mass from the rotation velocity, $v^2 r/G_N$, obtained from the farthest point to the centre of the galaxy. However, there seems to be a correlation with the centripetal acceleration $a = v^2/r$, $M/L \propto 1/a$ for $a < 10^{-8} \text{ cm} \cdot \text{s}^{-2}$. This is only a rephrasing of the Tully–Fischer relation. Milgrom [96] noticed that a way to recover this relation was to modify the Newtonian theory of gravity in the small acceleration regime, thus its name MOND (modified Newtonian dynamics).

MOND can be formulated either as a modification of Newton's second law of motion, or as a modification of gravity. In the first version, the relation between force and acceleration, $F = ma$, is modified to

$$m a \mu \left(\frac{a}{a_0} \right) = F, \quad (7.205)$$

where the constant a_0 has to be of the order $a_0 \sim 10^{-10} \text{ m/s}^2$ and where the fitting function μ should satisfy the asymptotic conditions

$$\mu(x) = \begin{cases} x & x \ll 1, \\ 1 & x \gg 1 \end{cases} \quad (7.206)$$

In the strong acceleration regimes, Newton's second law of motion is indeed recovered. In the second formulation, we assume that the gravitational acceleration, g , is related to the Newtonian gravitational acceleration, g_N , by

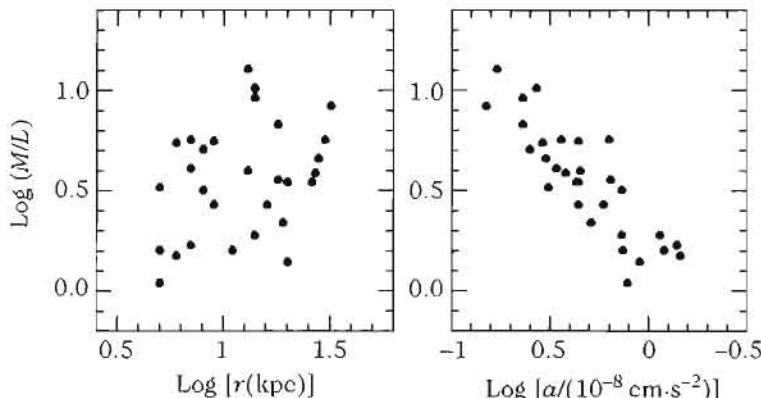


Fig. 7.25 Newtonian mass to luminosity ratio for spiral galaxies of Ursa Major as a function of the size of the galaxy (left) and as a function of the Newtonian acceleration (right). From Ref. [97].

$$g\mu \left(\frac{g}{a_0} \right) = g_N. \quad (7.207)$$

This corresponds to modifying the Poisson equation according to

$$\nabla \cdot \left[\mu \left(\frac{|\nabla \Phi|}{a_0} \right) \nabla \Phi \right] = 4\pi G_N \rho. \quad (7.208)$$

Even though there are differences in principle and in practice between these two formulations, they lead to the same predictions in the small-acceleration regime. In particular, the effective gravitational field becomes $g = \sqrt{g_N a_0}$ and the Keplerian velocity of a test body orbiting around a body of mass M is

$$v^4 = G_N M a_0. \quad (7.209)$$

The constant a_0 must be the same for all galaxies. It is associated with a characteristic scale $\ell_0 = c^2/a_0 \sim 9 \times 10^{26}$ m that, surprisingly, is of the same order of magnitude as c/H_0 . The transition scale between both kinds of regimes depends on the mass of the galaxy and is typically given by

$$\ell_* = \sqrt{\frac{\ell_0 G_N M}{c^2}} = 10^{20} \left(\frac{M}{10^{11} M_\odot} \right)^{1/2} \left(\frac{a_0}{10^{-10} \text{m.s}^{-2}} \right)^{-1/2} \text{m}. \quad (7.210)$$

7.2.5.12 Some predictions

In the MOND context, the existence of flat rotation curves has an absolute character. In particular, this hypothesis can be disproved by the observation of isolated galaxy rotation curves that would decrease in a Keplerian way at large distances, thus merely indicating the end of the dark-matter halo.

This model also has a number of predictions that we briefly summarize (see Ref. [97] for the current observational status).

- The Tully–Fisher relation $L \propto v_\infty^\alpha$ is absolute, having an exponent $\alpha = 4.0$, and gives a relation between the baryonic mass and the asymptotic velocity of the galaxy, namely

$$\ln L = 4 \ln v_\infty - \ln \left(G_N a_0 \left\langle \frac{M_b}{L} \right\rangle \right), \quad (7.211)$$

in agreement with the observed slope, $\alpha = 3.9 \pm 0.2$.

- One can define a critical surface density $\Sigma_m = a_0/G_N$ delimiting between two regimes. For $\Sigma > \Sigma_m$ (such as for galaxies with high surface density), there is little difference between the dynamical and visible mass, i.e. the galaxy does not seem to include any dark matter. For $\Sigma < \Sigma_m$ (such as for galaxies with a low surface density), the acceleration is weak and we are in the MOND regime. For these galaxies, there must be a large difference between the dynamical and visible mass; it seems to be dominated by dark matter. This is the case observationally.
- $\Sigma_m/2\pi \sim 140 M_\odot/\text{pc}^2$ must be the limiting value of the galaxy surface density supported by the rotation. This observed limit appears naturally in the MOND context.
- The rotation curve of a galaxy can be obtained uniquely from the knowledge of the baryonic matter distribution.

These four predictions have been verified observationally and in particular, all of the galaxy rotation curves seem to be reproducible with a unique parameter, the ratio M_b/L between the baryonic and luminous mass (which is in fact almost a constant).

Concerning lensing, it is clear from the form (7.204) of the potential that, at distances larger than $r \sim \sqrt{G_N M/a_0}$, the deflection angle is

$$\alpha = \frac{2\pi}{c^2} \sqrt{G_N M a_0},$$

that is exactly the same as in the case of a dark-matter halo when gravity is assumed to be described by general relativity.

7.2.5.13 Some problems

Because of the simplicity of its formulation and of its predictions, the MOND model remains a seductive alternative to the existence of dark matter. Nevertheless, several problems are present.

- From a theoretical point of view, there are no satisfactory field theories leading to MOND in the weak-field limit, despite a recent proposal of TeVeS theories [98]. In particular such a theory seems to require an acceleration a_0 and requires us to fix a reference frame. The existence of such a preferred frame is very constrained by tests of gravity in the Solar System. See also Ref. [99] for an in depth discussion of the difficulties to construct field theory versions of MOND.
- Galaxy clusters have a surface density that is important enough to be in the Newtonian regime for which MOND does not differ from Newtonian theory. But strong gravitational lensing indicates the need for dark matter in this regime.

- The required value of a_0 needed to explain observations of rich clusters is twice as big as that required to explain the galaxy rotation curves. It seems that the dynamics of elliptical galaxies requires $a_0 \sim 10^{-9} \text{ m} \cdot \text{s}^{-1}$.

Table 7.6 summarizes the comparative predictions from the MOND and SCDM models on a rating scale from 0 to 3. This summary suggests that the construction of a field theory capable of including MOND represents an important challenge. Moreover, construction of tests of the theory of gravity at cosmological scales is also a priority in order to be sure of the conclusions on the properties of dark matter.

Table 7.6 Summary of comparative predictions from MOND and SCDM models. Crosses represent difficulties and question marks are questions that cannot be rigorously addressed in the current framework.

| Test/predictions | MOND | SCDM |
|--|------|------|
| Rotation curves of spiral galaxies | 3 | 1 × |
| Tully–Fischer relation | 3 | × |
| Rotation curves from visible matter | 3 | × |
| Properties of elliptical galaxies | 1? | 1 |
| Temperature profile of clusters | × | 1 |
| Cosmic microwave background | ? | 3 |
| Large-scale structures and nucleosynthesis | ? | 3 |
| Correct light deflection | ×? | 2 |
| Satisfying theoretical framework | × | 2 |

7.2.5.14 Dark matter vs. modified general relativity

The former discussion concerning MOND relies on a Newtonian version of the law of gravity. Thus, a complete theory leading to MOND should arise from a modification of general relativity⁸ since the latter weak field limit is indeed Newton gravity. Various proposals have been investigated in the literature and we refer the reader to Ref. [99] for an extensive review.

In general relativity, the metric propagates as a pure, massless spin-2 field and matter is universally coupled to this metric. To extend general relativity, one must introduce new degrees of freedom into the theory, besides the graviton and the various matter fields entering the standard model of particle physics. This already renders the distinction with dark-matter models actually rather subtle.

The principal difference is that in the dark-matter model the amount of dark matter is imposed by initial conditions and its clustering generates gravitational wells in which baryonic matter fall to form large-scale structures. In a modified general relativity model, the baryonic matter generates by itself an effective dark-matter halo, scaling as $M_{\text{DM}} \propto \sqrt{M_{\text{baryon}}}$. Such a halo may thus merely be an artefact of the way

⁸We do not use the term ‘modified gravity’ since we define gravity as the only long-range force that cannot be screened. What could be modified is only the theory that describes gravity, which we assumed to be general relativity.

we interpret the gravitational field of baryonic matter alone at large distances. But it may also be a real dark-matter halo, made of the new (gravitational) fields that generates itself a Newtonian potential.

Whatever the class of models, the theory has to involve new fields, let us call them generically ψ . There is a stress-energy tensor associated to these fields and it will enter the 'Einstein' equations. We have various possibilities:

- $|T_{\mu\nu}^{(\psi)}| \gg |T_{\mu\nu}^{(\text{mat})}|$ and ψ is minimally coupled to the metric, in which case we have a dark-matter model. In the Newtonian regime, it is equivalent to adding a new component of matter and to assume the validity of the Poisson equation sourced by this new component.
- $|T_{\mu\nu}^{(\psi)}| \ll |T_{\mu\nu}^{(\text{mat})}|$, in which case the contribution of the new field to the energy density is negligible. This new field must then be responsible for a new interaction. We can say that we have a modified general relativity model. In the Newtonian regime, it modifies the Poisson equation and assumes that the energy-density is that of the baryons.
- $|T_{\mu\nu}^{(\psi)}| \gg |T_{\mu\nu}^{(\text{mat})}|$ and the new field is responsible for a new interaction. It is then difficult to decide whether we have a dark-matter model or a modified general relativity model.

Whatever the class of models, it is important to realize that we have to add new degrees of freedom in the currently accepted theories of physics. This ought to be the most fundamental conclusion that can be drawn from observational lensing.

References

- [1] A.O. PETERS, H. LEVINE and J. WAMBSGANSS, *Singularity theory and gravitational lensing*, Birkhauser, Boston, 1991.
- [2] P. SCHNEIDER, J. EHLLERS and E.E. FALCO, *Gravitational lenses*, Springer, Berlin, 1992.
- [3] J. WAMBSGANSS, 'Gravitational lensing in Astronomy', *Living Reviews Rel.* **1**, 1, 1998.
- [4] R. NARAYAN and M. BARTELmann, 'Gravitational lensing', in *Formation of structure in the Universe*, A. Dekel and J. Ostriker (eds.), Cambridge University Press, 1999, pp. 360.
- [5] W.L. BURKE, 'Multiple gravitational imaging by distributed masses', *Astrophys. J.* **244**, L1, 1981.
- [6] R.D. BLANDFORD *et al.*, 'Gravitational lens optics', *Nature* **245**, 824, 1989.
- [7] B. PACZYŃSKI, 'Gravitational microlensing in the local group', *Ann. Rev. Astron. Astrophys.* **34**, 419, 1996.
- [8] E. ROULET and S. MOLLERACH, 'Microlensing', *Phys. Rep.* **279**, 67, 1997.
- [9] B. FORT and Y. MELLIER, 'Arc(let)s in clusters of galaxies', *Astron. Astrophys. Review* **5**, 239, 1994.
- [10] F. COURBIN, 'Quasar lensing', *Lect. Notes Phys.* **608**, 1, 2002.
- [11] M. BARTELmann and P. SCHNEIDER, 'Weak gravitational lensing', *Phys. Rep.* **340**, 291, 2001.
- [12] Y. MELLIER, 'Probing the Universe with weak lensing', *Annu. Rev. Astron. Astrophys.* **37**, 127, 1999.
- [13] C. ALCOCK *et al.*, 'Possible gravitational microlensing of a star in the large Magellanic cloud', *Nature* **365**, 621, 1993.
- [14] E. AUBOURG *et al.*, 'Evidence for gravitational microlensing by dark object in the galactic halo', *Nature* **365**, 623, 1993.
- [15] A. MILSZTAJN, 'The galactic halo from microlensing', *Space Sci. Rev.* **100**, 103, 2002.
- [16] D. WALSH, R.F. CARSWELL and R.J. WEYMAN, '0957+561 A, B: twin quasistar objects or gravitational lens?', *Nature* **279**, 384, 1979.
- [17] M.J. IRWIN *et al.*, 'Photometric variations in the Q2237+035 system: first detection of a microlensing event', *Astrophys. J.* **98**, 1989, 1989.
- [18] C.S. KOCHANEK, 'The analysis of gravitational lens survey. II. Maximum likelihood models and singular potentials', *Astrophys. J.* **419**, 12, 1993.
- [19] S. MAO and C.K. KOCHANEK, 'Limits on galaxy evolution', *Month. Not. R. Astron. Soc.* **268**, 569, 1994.
- [20] R. NARAYAN and S. WHITE, 'Gravitational lensing in cold dark-matter Universe', *Month. Not. R. Astron. Soc.* **231**, 97, 1988.

- [21] E.L. TURNER, 'Gravitational lensing limits on the cosmological constant in flat Universe', *Astrophys. J.* **365**, L43, 1990.
- [22] C.K. KOCHANEK, 'Is there a cosmological constant?', *Astrophys. J.* **466**, 638, 1996.
- [23] M. IM *et al.*, 'A measurement of the cosmological constant using elliptical galaxies as strong gravitational lenses', *Astrophys. J.* **497**, 457, 1997.
- [24] S. REFSDAL, 'The gravitational lens effect', *Month. Not. R. Astron. Soc.* **128**, 295, 1964.
- [25] C. KOCHANEK, 'What do gravitational lens delays measure?', *Astrophys. J.* **578**, 25, 2002.
- [26] N. GROGIN and R. NARAYAN, 'A new model of the gravitational lens 0957+561 and a limit on the Hubble constant', *Astrophys. J.* **464**, 92, 1996.
- [27] T. KUNDIĆ *et al.*, 'A robust determination of the time delay in 0957+561A,B and a measurement of the global value of the Hubble constant', *Astrophys. J.* **482**, 75, 1996.
- [28] E.E. FALCO *et al.*, 'An estimate of H_0 from Keck spectroscopy of gravitational lens system 0957+561', *Astrophys. J.* **484**, 70, 1997.
- [29] G. SOUCAIL *et al.*, 'A blue ring-like structure in the centre of the A370 cluster of galaxies', *Astron. Astrophys.* **172**, L14, 1987.
- [30] R. LYNGS and V. PETROSIAN, 'Giant luminous arcs in Galaxy clusters', *BAAS* **18**, 1014, 1986.
- [31] J.-P. KNEIB *et al.*, 'Redshift survey up to $b_J = 27$: distance of gravitational arclets behind Abell 370', *Astron. Astrophys.* **286**, 701, 1994.
- [32] B. FORT *et al.*, 'Distribution of galaxies at large redshift and cosmological parameters from the magnification bias in CL0024+1654', *Astron. Astrophys.* **321**, 353, 1997.
- [33] R. PELLO *et al.*, 'ISAAC/VLT observations of a lensed galaxy at $z = 10.0$ ', *Astron. Astrophys.* **416**, L35, 2004.
- [34] C. SCHIMD, J.-P. UZAN and A. RIAZUELO, 'Weak lensing in scalar tensor theories of gravity', *Phys. Rev. D* **71**, 083512, 2005.
- [35] W. HU, 'Weak lensing of the CMB: a harmonic approach', *Phys. Rev. D* **62**, 043007, 2000.
- [36] N. KAISER and G. SQUIRES, 'Mapping the dark matter with weak gravitational lensing' *Astrophys. J.* **404**, 441, 1993.
- [37] P. SCHNEIDER *et al.*, 'B-modes in cosmic shear from source redshift clustering', *Astron. Astrophys.* **389**, 729, 2002.
- [38] D. WITTMAN, 'Weak Lensing', *Lec. Notes Phys.* **608**, 55, 2002.
- [39] L. VAN WAERBEKE *et al.*, 'Detection of correlated galaxy ellipticities from CFHT data: first evidence for gravitational lensing by large-scale structures', *Astron. Astrophys.* **358**, 30, 2000.
- [40] D. BACON *et al.*, 'Detection of weak gravitational lensing by large-scale structure', *Month. Not. R. Astron. Soc.* **318**, 625, 2000.
- [41] N. KAISER *et al.*, 'Large-scale cosmic shear measurements', [[astro-ph/0003338](#)].
- [42] D. WITTMAN *et al.*, 'Detection of weak gravitational lensing distortions of distant galaxies by cosmic dark matter at large scales', *Nature* **405**, 143, 2000.

- [43] L. VAN WAERBEKE *et al.*, 'Cosmic shear statistics and cosmology', *Astron. Astrophys.* **375**, 757, 2001.
- [44] L. FU *et al.*, 'Very weak lensing in the CFHTLS wide: cosmology from cosmic shear in the linear regime', *Astron. Astrophys.* **479**, 9, 2008.
- [45] U. SELJAK, 'Gravitational lensing effect on CMB anisotropies: a power approach', *Astrophys. J.* **463**, 1, 1996; W. HU, 'Weak lensing on the CMB: a harmonic approach', *Phys. Rev. D* **62**, 043007, 2000; A. CHALLINOR and A. LEWIS, 'Lensed CMB power spectra from all-sky correlation functions', [[astro-ph/0502425](#)].
- [46] M. ZALDARRIAGA and U. SELJAK, 'Gravitational lensing effect on cosmic microwave background', *Phys. Rev. D* **58**, 023003, 1998.
- [47] F. COMBES, 'Properties of dark-matter haloes', *New Astron. Rev.* **46**, 755, 2002.
- [48] R. GENZEL *et al.*, 'Near infrared flares from accreting gas around the supermassive black hole at the galactic centre', *Nature* **425**, 934, 2003.
- [49] Y. SOFUE *et al.*, 'Central rotation curves of spiral galaxies', *Astrophys. J.* **523**, 136, 1999.
- [50] S. McGAUGHEY *et al.*, 'The baryonic Tully–Fischer relation', *Astrophys. J.* **533**, L99, 2000.
- [51] A. BOESMA, '21-cm line studies of spiral galaxies. I.', *Astron. J.* **86**, 1791, 1981.
- [52] H. HOEKSTRA *et al.*, 'On the apparent coupling of neutral hydrogen and dark matter in spiral galaxies', *Month. Not. R. Astron. Soc.* **323**, 453, 2001.
- [53] D. PFENNIGER and Y. REVAZ, 'The baryonic Tully–Fischer relation revisited', *Astron. Astrophys.* **431**, 511, 2005.
- [54] D. PFENNIGER and F. COMBES, 'Is dark matter in spiral galaxies cold gas?' *Astron. Astrophys.* **285**, 94, 1994.
- [55] M. PERSIC, P. SALUCCI and F. STEL, 'The universal rotation curve of spiral galaxies-I. The dark matter connection', *Month. Not. R. Astron. Soc.* **281**, 27, 1996.
- [56] F. DONATO, G. GENTILE and P. SALUCCI, 'Cores of dark-matter halos correlate with stellar scale lengths', *Month. Not. R. Astron. Soc.* **353**, 17, 2004.
- [57] M. BERNARDI *et al.*, 'Early type galaxies in the Sloan Digital Sky Survey: the fundamental plane', *Astrophys. J.* **125** (2003) 1866.
- [58] S.G. DJORGOVSKI and M. DAVIS, 'Fundamental properties of elliptical galaxies', *Astrophys. J.* **313**, 59, 1987; A. DRESSLER *et al.*, 'Spectroscopy and photometry of elliptical galaxies', *Astrophys. J.* **313**, 42, 1987.
- [59] D. MERRITT, 'A universal density profile for dark and luminous matter', *Astrophys. J.* **624**, L85, 2005.
- [60] F. ZWICKY, 'On the masses of nebulae and of clusters of nebulae', *Astrophys. J.* **86**, 217, 1937.
- [61] N. BAHCALL, 'Clusters and superclusters of galaxies', [[astro-ph/9611148](#)].
- [62] H. HOEKSTRA *et al.*, 'Properties of galaxy dark-matter halos from weak lensing', *Astrophys. J.* **606**, 67, 2004.
- [63] F. PRADA *et al.*, 'Observing the dark-matter density profile of isolated galaxies', *Astrophys. J.* **598**, 260, 2003.
- [64] P. COLÍN *et al.*, 'Substructure and halo density profiles in a warm dark-matter cosmology', *Astrophys. J.* **542**, 622, 2000.

- [65] B.N. SPERGEL and P.J. STEINHARDT, 'Observational evidence for self-interacting cold dark matter', *Phys. Rev. Lett.* **84**, 3760, 2000.
- [66] J. GOODMAN, 'Repulsive dark matter', *New Astron.* **5**, 103, 2000.
- [67] P.J.E. PEEBLES and A. VILENKIN, 'Noninteracting dark matter', *Phys. Rev. D* **60**, 103506, 1999.
- [68] W. HU *et al.*, 'Cold and fuzzy dark matter', *Phys. Rev. Lett.* **85**, 1158, 2000.
- [69] L. KAPLINGHAT *et al.*, 'Annihilating dark matter', *Phys. Rev. Lett.* **85**, 3335, 2000.
- [70] R. CEN, 'Why are there dwarf spheroidal galaxies?', *Astrophys. J.* **456**, L77, 2001.
- [71] M. KAMIONKOWSKI and A. LIDDLE, 'The dearth of halo of dwarf galaxies: is there power on short scales?', *Phys. Rev. Lett.* **84**, 4525, 2000.
- [72] J.P. OSTRIKER and P.J. STEINHARDT, 'New light on dark matter', *Science* **300**, 1909, 2003.
- [73] G. BERTONE, D. HOOPER and J. SILK, 'Particle dark matter: evidence, candidates and constraints', *Phys. Rep.* **405**, 279, 2005.
- [74] P. GONDOLO, 'Non-baryonic dark matter', *NATO Sci. Ser. II* **187**, 279, 2005.
- [75] J. ELLIS, 'Partical candidates for dark matter', *Phys. Scripta* **T85**, 221, 2000.
- [76] L. BERSTRÖM, 'Non baryonic dark matter', *Rept. Prog. Phys.* **63**, 793, 2000.
- [77] A.D. DOLGOV 'Neutrinos in cosmology', *Phys. Rep.* **370**, 333, 2002.
- [78] J. LESGOURGUES and S. PASTOR 'Massive neutrinos and cosmology', *Phys. Rep.* **49**, 307, 2006.
- [79] S. TREMAINE and J.S. GUNN, 'Dynamical role of light neutral leptons in cosmology', *Phys. Rev. Lett.* **42**, 407, 1979.
- [80] S. DODELSON and L.M. WIDROW, 'Sterile neutrinos as dark matter', *Phys. Rev. Lett.* **72**, 17, 1994.
- [81] D. BAILIN and A. LOVE, *Introduction to gauge field theory*, Bristol: IOP, 1993.
- [82] R. PECCEI and H.R. QUINN, 'CP conservation in the presence of instantons', *Phys. Rev. Lett.* **38**, 1440, 1977.
- [83] T. MOROI *et al.*, 'Cosmological constraints on the light stable gravitino', *Phys. Lett. B* **303**, 289, 1993.
- [84] G. JUNGMAN *et al.*, 'Supersymmetric dark matter', *Phys. Rep.* **267**, 195, 1996.
- [85] B.W. LEE and S. WEINBERG, 'Cosmological lower bound on heavy-neutrino masses', *Phys. Rev. Lett.* **39**, 165, 1977.
- [86] D. CHUNG, E. KOLB and A. RIOTTO, 'Nonthermal supermassive dark matter', *Phys. Rev. Lett.* **81**, 4048, 1998; D. CHUNG, E. KOLB and A. RIOTTO, 'Super-heavy dark matter', *Phys. Rev. D* **59**, 023501, 1999.
- [87] V. KUZMIN and I. TKACHEV, 'Ultra-high energy cosmic rays, superheavy long-living particles, and matter creation after inflation', *JETP Lett.* **68**, 271, 1998.
- [88] C. BOEHM and P. FAYET, 'Scalar dark-matter candidates', *Nucl. Phys. B* **683**, 219, 2004.
- [89] N. ARKANI-HAMED *et al.*, 'Electroweak symmetry breaking from dimensional deconstruction', *Phys. Lett. B* **513**, 232, 2001.
- [90] H.C. CHENG and I. LOW, 'TeV symmetry breaking and little hierarchy problem', *JHEP* **0309**, 051, 2003.

- [91] E.W. KOLB and R. SLANSKY, 'Dimensional reduction in the early Universe: where have the massive particles gone?', *Phys. Lett. B* **135**, 378, 1984.
- [92] B. CARTER, 'Recent developments in vorton theory', *Int. J. Theor. Phys.* **36**, 2451, 1997.
- [93] G. CHARDIN, *Dark matter: direct detection*, in *The primordial Universe*, Ecole des Houches, Session LXXI, P. Binetruy et al. Eds., EDP Sciences, 2000.
- [94] L. ROSZKOWSKI, 'Particle dark matter - a theorist's prospective', *Pramana* **62**, 389, 2004.
- [95] K. OLIVE, 'Searching for supersymmetric dark matter', [hep-ph/0208092].
- [96] M. MILGROM, 'A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis', *Astrophys. J.* **270**, 365, 1983.
- [97] R.H. SANDERS and S.S. McGAUGHEY, 'Modified Newtonian dynamics as an alternative to dark matter', *Ann. Rev. Astron. Astrophys.* **40**, 263, 2002.
- [98] J. BEKENSTEIN, 'Relativistic gravitation theory for modified Newtonian dynamics paradigm', *Phys. Rev. D* **70**, 083509, 2004.
- [99] J.-P. BRUNETON and G. ESPOSITO-FARÈSE, 'Field theoretical formulations of MOND-like gravity', *Phys. Rev. D* **76**, 124012, 2007.

8

Inflation

Inflation is a postulated period of accelerated expansion occurring before the radiation era of the standard Big-Bang model. It was initially proposed as a tentative solution of the problems of the hot Big-Bang model (Chapter 4). In particular, it provides a simple explanation for the homogeneity and flatness of our Universe.

Before the first models of inflation were proposed, precursor works appeared as early as 1965 by Gliner [1], who postulated a phase of exponential expansion. In 1978, Englert, Brout and Gunzig [2], in an attempt to resolve the primordial singularity problem and to introduce the particles and the entropy contained in the Universe, proposed a ‘fireball’ hypothesis, whereby the Universe itself would appear through a quantum effect in a state of negative pressure subject to a phase of exponential expansion.

Starobinsky [3], in 1979, used quantum-gravity ideas to formulate the first semi-realistic rigorous model of an inflation era, although he did not aim to solve the cosmological problems. A simpler model, with transparent physical motivations, was then proposed by Guth [4] in 1981; this model, now called ‘old inflation’, was the first to use inflation as a means of solving cosmological problems. It was soon followed by Linde’s ‘new inflation’ proposal [5].

The formulation by Mukhanov and Chibisov of the theory of cosmological perturbations during new inflation, which links the origin of the large-scale structures of the Universe to quantum fluctuations, quickly followed [6, 7]. All these important works were achieved in the early 1980’s, which can thus be considered as the ‘date of birth’ of inflation.

The first models of inflation were actually only incomplete modifications of the Big-Bang theory as they assumed that the Universe was in a state of thermal equilibrium, homogeneous on large enough scales before the period of inflation. This problem was resolved by Linde with the proposition of chaotic inflation [8]. In this model, inflation can start from a Planckian density even if the Universe is not in equilibrium. A new picture of the Universe then appears. The homogeneity and isotropy of our observable Universe would be only local properties, while the Universe is very inhomogeneous on very large scales, with a fractal-like structure.

This chapter describes the first models of inflation, Section 8.1, and explains in general terms how inflation solves the hot Big-Bang problems. We then address, in Section 8.2, the dynamics of single-field inflation and discuss the slow-roll approximation and the chaotic model of inflation. The generation of initial density perturbations during the inflationary epoch is detailed in Section 8.3, 8.4 and 8.5. To finish, we describe the reheating phase in Section 8.6, which connects inflation to the hot Big-Bang

model, the mechanism of eternal inflation, Section 8.7, which can address the question of initial conditions of the Universe, and various possible extensions, Section 8.8.

Numerous complementary studies can be found in Refs. [9–11].

8.1 Genesis of the paradigm

8.1.1 Original motivations

8.1.1.1 Guth's argument

The flatness problem can be reformulated by noting that

$$\frac{\Omega_K}{\Omega} = -\frac{3K}{8\pi G_N \rho a^2},$$

so that (4.20) and (4.21) lead to the conclusion that

$$\left| \frac{\Omega_K}{\Omega} \right| \sim \frac{10^{43}}{S_0^{2/3}} \left(\frac{t}{1 \text{ s}} \right) \sim \frac{10^{37}}{S_0^{2/3}} \left(\frac{1 \text{ GeV}}{T} \right)^2, \quad (8.1)$$

where $S_0 = s_0 H_0^{-3}$ is the entropy contained in the Hubble radius today. From (4.140), it is of order $S_0 \sim 10^{87}$. Such a large value of S_0 implies that the Universe should have been very flat in the past. For instance, the curvature parameter must typically be of order $\Omega_K \sim 10^{-16}$ at the time of nucleosynthesis.

The flatness problem is therefore related to the fact that the entropy in a comoving volume is conserved. One can then imagine, as proposed by Guth [4], that it can be resolved if the cosmic expansion is non-adiabatic during a period of time $[t_i, t_f]$ during which

$$S_f = Z^3 S_i, \quad (8.2)$$

where Z is a numerical factor to be determined. In Guth's proposal, this entropy production occurs at around $T_{\text{GUT}} \sim 10^{17}$ GeV, the characteristic scale of the Grand Unified Theory (GUT) symmetry breaking (see Chapter 9). From (8.1), this implies that Z should be of order $Z \sim 10^{28}$ in order for the curvature parameter to be of order unity today, $\Omega_{K0} \sim \mathcal{O}(1)$.

8.1.1.2 A phase of acceleration

The Friedmann equations (3.25) indicate that $(1 - \Omega^{-1})\rho a^2 = 3K/8\pi G_N$ is constant during cosmic evolution. We infer that

$$(1 - \Omega_i^{-1})\rho_i a_i^2 = (1 - \Omega_f^{-1})\rho_f a_f^2.$$

Assuming that the subsequent evolution of the Universe ($t > t_f$) is described by the standard Big-Bang model, and is thus adiabatic, we conclude that

$$(1 - \Omega_i^{-1})\rho_i a_i^2 \sim 10^{-56} (1 - \Omega_0^{-1})\rho_f a_f^2.$$

The flatness problem can thus be resolved if $\rho_f a_f^2 \gg \rho_i a_i^2$. Moreover, since $H^2 a^2 - 8\pi G_N \rho a^2 / 3$ is constant in time, we have that

$$\dot{a}_f > \dot{a}_i. \quad (8.3)$$

So, a necessary condition for inflation is that the expansion of the Universe be accelerated during the period $[t_i, t_f]$, which implies that during this period the strong energy condition must be violated [see (3.26)]:

$$\ddot{a}(t) > 0 \iff \rho + 3P < 0. \quad (8.4)$$

8.1.2 Resolution of the Big-Bang problems

Inflation is therefore defined as a phase of accelerated expansion of the Universe. This property is sufficient to show that the Big-Bang problems can be resolved if this phase lasts sufficiently long.

To quantify the length of the inflationary period, we define the quantity

$$N \equiv \ln \left(\frac{a_f}{a_i} \right), \quad (8.5)$$

where a_i and a_f are the values of the scale factor at the beginning and at the end of inflation. This number measures the growth in the scale factor during the accelerating phase, expressed in units of the exponential of base ‘e’, which is the reason why it is usually called the *number of ‘e-folds’*¹.

8.1.2.1 Flatness problem

The dynamical analysis of the Friedmann–Lemaître equations (Chapter 3) has shown (3.54) that, when $\Omega_\Lambda = 0$,

$$\frac{d\Omega_K}{d \ln a} = (1 + 3w)\Omega_K(1 - \Omega_K).$$

When $-1 < w < -1/3$, the stable fixed point of this dynamical system is on the $\Omega_K = 0$ axis (see Figs. 3.4–3.6). This attractor becomes unstable when w becomes larger than $-1/3$, that is in the matter and radiation eras. So, the flatness problem will be resolved if the attraction toward $\Omega_K = 0$ during the period of inflation is sufficient to compensate its subsequent drift during the hot Big-Bang, i.e. if inflation has lasted sufficiently long.

To give an estimate of the required minimum number of e-folds of inflation, note that, if we assume H to be almost constant during inflation then

$$\left| \frac{\Omega_K(t_f)}{\Omega_K(t_i)} \right| = \left(\frac{a_f}{a_i} \right)^{-2} = e^{-2N}.$$

In order to have $|\Omega_K(t_f)| \lesssim 10^{-60}$ and $\Omega_K(t_i) \sim \mathcal{O}(1)$, we thus need

$$N \gtrsim 70. \quad (8.6)$$

¹Instead of the number of e-folds, the international nomenclature uses here the neper (symbol Np) which is used to express, for instance, the value of logarithmic quantities such as power ratios or acoustic pressure. The neper is consistent with the International System (IS) but has not yet been adopted by the General Conference as an IS unit. The neper is analogous to the decibel, with the convention that $1 \text{ Np} = 20/(\ln 10) \text{ dB} = 8.686 \text{ dB}$, which expresses the use of a natural logarithmic basis rather than decimal for the dB. In what follows, we will use the term ‘e-fold’ as it is the one found in the literature.

8.1.2.2 Horizon problem

The hypothesis (8.4) implies that the comoving Hubble radius, $\mathcal{H}^{-1} = (aH)^{-1}$, decreases in time

$$\frac{d}{dt}(aH)^{-1} < 0. \quad (8.7)$$

Two points in causal contact at the beginning of inflation can thus be separated by a distance larger than the Hubble radius at the end of inflation (Fig. 8.1). These points are still causally connected, but can seem to be causally disconnected if the inflationary period is omitted. So, inflation allows for the entire *observable* Universe to emerge out of the same causal region before the onset of inflation.

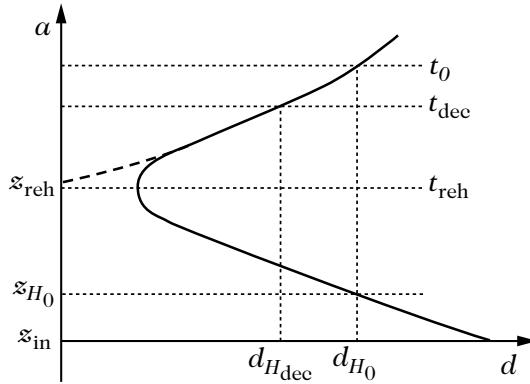


Fig. 8.1 Evolution of the comoving Hubble radius as a function of the scale factor of the Universe assuming a period of inflation (plain line). A given comoving scale (vertical line) exits the Hubble radius during inflation and enters back into the Hubble radius during the hot Big-Bang. All observable scales can thus have been in causal contact at the beginning of inflation. This is not the case without inflation (dashed line) since the comoving Hubble radius is then strictly growing.

For the horizon problem to be solved *today*, a causal region with scale factor a_i must have at least the size of the current observable Universe, which implies that

$$e^N \sim \frac{a_f}{a_0} \frac{D_H(t_0)}{D_H(t_f)} \sim \frac{T_0}{T_f} \frac{D_H(t_0)}{D_H(t_f)}.$$

Assuming that inflation ends at the Grand Unification scale (see Chapter 9), $T_f \sim 10^{16}$ GeV then $D_H(t_f) \sim M_P/T_f^2 \sim 10^{-13}$ GeV $^{-1}$. Since $T_0 \sim 10^{-4}$ eV and $D_H(t_0) \sim H_0^{-1} \sim 10^{41}$ GeV $^{-1}$, we infer that

$$N \gtrsim 57. \quad (8.8)$$

This value can be smaller if the energy scale at the end of inflation is lower.

8.1.2.3 Conclusions

A primordial period of accelerated expansion allows for the resolution of the hot Big-Bang problems if it lasts sufficiently long. We have estimated that the number of e-folds should typically be at least of the order

$$N \gtrsim 50 - 70, \quad (8.9)$$

depending on what the energy scale is at the end of inflation. It can be shown that with such a value, the other problems (monopoles, homogeneity, entropy) are also resolved.

It remains to build physical models leading to such an inflationary phase and to understand how it connects with the hot Big-Bang model.

8.1.3 First models of inflation

Every inflationary model relies on one or several scalar fields, called ‘inflaton’.

8.1.3.1 ‘Old’ inflation

Models of old inflation rely on a first-order phase-transition mechanism (Fig. 8.2). A scalar field is trapped in a local minimum of its potential, thus imposing a constant energy density, $V(\varphi_i)$, equivalent to the contribution of a cosmological constant. As long as the field remains in this configuration, the evolution of the Universe is exponential and the Universe can be described by a de Sitter space-time.

This configuration is metastable and the field can tunnel to its global minimum, $V(\varphi_f) = 0$, hence creating bubbles of true vacuum, which correspond to a non-inflationary Universe (see Fig. 8.2).

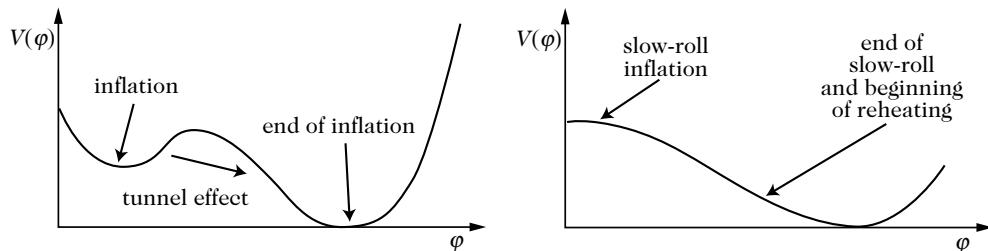


Fig. 8.2 The model of old inflation (left) is based on a first order phase transition from a local to a true minimum of the potential. The false vacuum is metastable and the field can tunnel to the true vacuum. In the models of new inflation (right) a scalar field slowly relaxes towards its vacuum.

One of the problems with these models arises from the properties of the de Sitter space (Section 8.1.4) that can be considered to be expanding, contracting or static space depending on the choice of coordinates (see below). In the metastable phase, there is a priori no preferred hypersurface and thus no preferred time-like direction to choose a slicing. Without such a direction, the phase transition can occur on any hypersurface. Each bubble of true vacuum will thus have very different physical properties,

e.g. spatial curvature, so that the Universe is expected to be very inhomogeneous on large scales. This led to the abandonment of this model [12].

8.1.3.2 New inflation

In models of *new inflation* [5, 13], a scalar field, φ , exits its false vacuum by slowly rolling towards its true vacuum (see Fig. 8.2). This phase can be preceded by a phase of old inflation. The slow-roll phase is essential as it is during this period that density fluctuations, which lead to the currently observed large-scale structures, are generated.

This model can only work if the potential has a very flat plateau around $\varphi = 0$, which is artificial. In most versions of this model, the inflaton cannot be in thermal equilibrium with other matter fields. The theory of cosmological phase transitions then does not apply.

No satisfying realization of this model has been proposed and it has been progressively abandoned.

8.1.4 Inflation as a de Sitter phase

In the limit of a frozen field, the inflationary phase is described by a de Sitter Universe. This space-time will be useful to understand and illustrate some mechanisms occurring during inflation. In realistic models of inflation, the dynamics of the inflaton induces a deviation of the space-time structure and that of a pure de Sitter space-time. As will also be seen, the limit where the field is completely frozen is usually singular.

8.1.4.1 Embedding in a five-dimensional space

The de Sitter space-time can be represented as a four-dimensional hyperboloid embedded in a five-dimensional Minkowski space,

$$-(z_0)^2 + (z_1)^2 + (z_2)^2 + (z_3)^2 + (z_4)^2 = H^{-2}. \quad (8.10)$$

This surface is invariant under five-dimensional Lorentz transformations, so that the de Sitter space is invariant under the 10-parameter symmetry group $O(4,1)$. It is therefore a maximally symmetric four-dimensional space, just as for Minkowski space-time, with curvature

$$R = 12H^2 = \text{const.} \quad (8.11)$$

Depending on the choice of the constant-time hypersurfaces, (see Fig. 8.3), the de Sitter space can take the form of a Friedmann–Lemaître space-time with flat, closed or open spatial sections, or a static form similar to the Schwarzschild space-time.

8.1.4.2 Euclidean slicing

The coordinate choice

$$z_0 = \frac{\sinh Ht}{H} + \frac{H}{2}e^{Ht}\mathbf{x}^2, \quad z_4 = \frac{\cosh Ht}{H} - \frac{H}{2}e^{Ht}\mathbf{x}^2, \quad z_i = e^{Ht}x_i, \quad (8.12)$$

leads to the expression for the induced metric

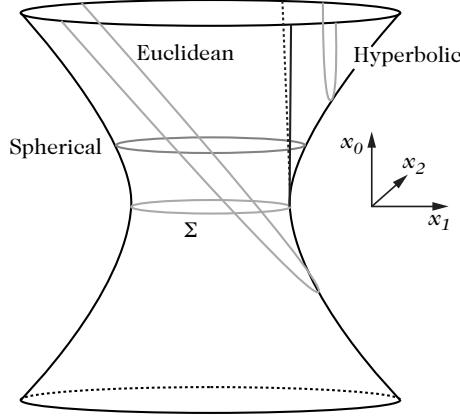


Fig. 8.3 Depending on the choice of the constant-time hypersurfaces, the four-dimensional surface (8.10) can take various forms.

$$ds^2 = -dt^2 + e^{2Ht} \delta_{ij} dx^i dx^j, \quad (8.13)$$

with Euclidean spatial sections (see Section 8.4.5 for details on the computation of an induced metric). In conformal time, this space takes the form

$$ds^2 = \frac{1}{(H\eta)^2} (-dt^2 + \delta_{ij} dx^i dx^j), \quad (8.14)$$

with

$$\eta = -\frac{1}{H} e^{-Ht}. \quad (8.15)$$

In this representation, η varies between $-\infty$ and 0^- , with the limit $\eta \rightarrow 0^-$ representing ‘time-like infinity’. However, it only covers half of the hyperboloid (8.10).

8.1.4.3 Spherical slicing

The choice of the system of coordinates (t, χ, θ, ϕ) defined by

$$\begin{aligned} z_0 &= \frac{\sinh Ht}{H}, & z_1 &= \frac{\cosh Ht}{H} \cos \chi, & z_2 &= \frac{\cosh Ht}{H} \sin \chi \cos \theta, \\ z_3 &= \frac{\cosh Ht}{H} \sin \chi \sin \theta \cos \phi, & z_4 &= \frac{\cosh Ht}{H} \sin \chi \sin \theta \sin \phi, \end{aligned} \quad (8.16)$$

can be used to rewrite the metric in the form of a Friedmann–Lemaître space-time with spherical spatial sections,

$$ds^2 = -dt^2 + \frac{\cosh^2 Ht}{H^2} (d\chi^2 + \sin^2 \chi d\Omega^2). \quad (8.17)$$

Unlike the Euclidean form (8.13), and the hyperbolic version, this slicing covers the entire de Sitter space-time and is thus geodesically complete.

8.1.4.4 Static slicing

To convince ourselves that de Sitter space-time is a static space-time, since it is maximally symmetric, let us derive its static form explicitly. With the choice of coordinates

$$\begin{aligned} z_0 &= \frac{\sinh Ht}{H} \sqrt{1 - r^2 H^2}, & z_1 &= \frac{\cosh Ht}{H} \sqrt{1 - r^2 H^2}, \\ z_2 &= r \cos \theta, & z_3 &= r \sin \theta \cos \phi, & z_4 &= r \sin \theta \sin \phi, \end{aligned} \quad (8.18)$$

where $0 \leq rH \leq 1$, the metric takes the Schwarzschild form

$$ds^2 = - (1 - r^2 H^2) dt^2 + \frac{dr}{(1 - r^2 H^2)} + r^2 d\Omega^2. \quad (8.19)$$

8.2 Dynamics of single-field inflation

The first models of inflation introduced two ingredients shared by a fair majority of inflation models: the existence of a phase during which the Universe is ‘close’ to a de Sitter space-time, and a slowly rolling scalar field that rules the expansion of the Universe.

8.2.1 Equations of evolution

The simplest models of inflation are constructed from a single scalar field, φ , evolving in a potential $V(\varphi)$ with an action given by

$$S = - \int \sqrt{-g} \left[\frac{1}{2} \partial_\mu \varphi \partial^\mu \varphi + V(\varphi) \right] d^4x, \quad (8.20)$$

so that its energy-momentum tensor takes the form

$$T_{\mu\nu} = \partial_\mu \varphi \partial_\nu \varphi - \left(\frac{1}{2} \partial_\alpha \varphi \partial^\alpha \varphi + V \right) g_{\mu\nu}. \quad (8.21)$$

The energy density and pressure of a homogeneous scalar field are, respectively, given by

$$\boxed{\rho_\varphi = \frac{\varphi'^2}{2a^2} + V(\varphi), \quad P_\varphi = \frac{\varphi'^2}{2a^2} - V(\varphi)}, \quad (8.22)$$

in conformal time, or by

$$\boxed{\rho_\varphi = \frac{\dot{\varphi}^2}{2} + V(\varphi), \quad P_\varphi = \frac{\dot{\varphi}^2}{2} - V(\varphi)}, \quad (8.23)$$

in cosmic time. These expressions show that $\rho_\varphi + 3P_\varphi = 2(\dot{\varphi}^2 - V)$. Since the Friedmann equations imply $\ddot{a}/a = -4\pi G_N(\rho + 3P)/3$ (see Chapter 3), the Universe can enter a phase of accelerated expansion as soon as $\dot{\varphi}^2 < V$. The expansion will be quasi-exponential if the scalar field is in a slow-roll regime, i.e. if $\dot{\varphi}^2 \ll V$.

8.2.1.1 *Equations in cosmic time*

The Friedmann and Klein–Gordon equation in cosmic time take the form (see Chapter 3 for their derivation)

$$H^2 = \frac{8\pi G_N}{3} \left(\frac{1}{2}\dot{\varphi}^2 + V \right) - \frac{K}{a^2}, \quad (8.24)$$

$$\frac{\ddot{a}}{a} = \frac{8\pi G_N}{3} (V - \dot{\varphi}^2), \quad (8.25)$$

$$\ddot{\varphi} + 3H\dot{\varphi} + V_{,\varphi} = 0. \quad (8.26)$$

Using $\dot{H} = \ddot{a}/a - H^2$, we obtain the relation

$$\dot{H} = -4\pi G_N \dot{\varphi}^2 + \frac{K}{a^2}, \quad (8.27)$$

which will be useful in what follows. As usual, only two of these equations are independent.

8.2.1.2 *Equations in conformal time*

These equations can be translated into conformal time. They then take the form

$$\mathcal{H}^2 = \frac{8\pi G_N}{3} \left(\frac{1}{2}\varphi'^2 + Va^2 \right) - K, \quad (8.28)$$

$$\mathcal{H}' = -\frac{8\pi G_N}{3} (\varphi'^2 - Va^2), \quad (8.29)$$

for the Friedmann equations, and

$$\varphi'' + 2\mathcal{H}\varphi' + a^2 V_{,\varphi} = 0, \quad (8.30)$$

for the Klein–Gordon equation.

8.2.1.3 *Hamilton–Jacobi equations*

As long as the evolution of the scalar field is monotonic, it can be useful to rewrite (8.24)–(8.26) using φ , instead of t or η , as a time variable. Assuming that $K = 0$, (8.27) takes the form

$$H_{,\varphi} \equiv \frac{dH(\varphi)}{d\varphi} = -4\pi G_N \dot{\varphi}, \quad (8.31)$$

which defines the notation $H_{,\varphi}$, so that (8.24) becomes

$$[H_{,\varphi}(\varphi)]^2 - 12\pi G_N H^2(\varphi) = -32\pi^2 G_N^2 V(\varphi). \quad (8.32)$$

Both (8.31) and (8.32) are strictly equivalent to (8.24)–(8.27). This formalism is very efficient to obtain exact solutions of the dynamics since the choice of a function $H(\varphi)$ directly determines the form of the required potential.

8.2.2 Slow-roll parameters

The slow-roll formalism is useful in reformulating most models of inflation that do not rely on first-order phase transitions (such as old inflation) in terms of the so-called *slow-roll parameters*. It makes it possible to compute the predictions of inflation without requiring a specification of the exact form of the inflaton's potential.

We assume that $K = 0$, since the curvature term decreases as a^{-2} and becomes negligible in comparison with the contribution from the scalar field, which has an almost constant energy density, if inflation lasts sufficiently long. This first implies that $\dot{H} < 0$ [see (8.27)], and thus H is always decreasing, and second that $|\dot{H}| < H^2$ if $\ddot{a} > 0$.

8.2.2.1 Principle

During inflation, the scalar field must be in a slow-roll regime. This implies that the scalar field varies little during the inflationary phase and thus satisfies

$$\dot{\varphi}^2 \ll V, \quad \ddot{\varphi} \ll 3H\dot{\varphi}. \quad (8.33)$$

Under these conditions, the evolution equations (8.24) and (8.26) reduce to

$$H^2 \simeq \frac{8\pi G_N}{3}V, \quad \dot{H} = -4\pi G_N \dot{\varphi}^2, \quad 3H\dot{\varphi} \simeq -V_{,\varphi}, \quad (8.34)$$

which implies

$$\frac{|\dot{H}|}{H^2} \ll \frac{3}{2}. \quad (8.35)$$

Moreover, since

$$\ddot{\varphi} \simeq -\frac{d}{dt} \left(\frac{V_{,\varphi}}{3H} \right) \simeq -\frac{V_{,\varphi\varphi}\dot{\varphi}}{3H} \simeq \frac{V_{,\varphi}V_{,\varphi\varphi}}{9H^2} \simeq \frac{1}{24\pi G_N} \frac{V_{,\varphi}V_{,\varphi\varphi}}{V},$$

we infer that the two conditions (8.33) imply

$$\left(\frac{V_{,\varphi}}{V} \right)^2 \ll 24\pi G_N, \quad \frac{|V_{,\varphi\varphi}|}{V} \ll 24\pi G_N. \quad (8.36)$$

So, in order for inflation to occur, the potential should be very flat. In this slow-roll regime, we can expand (8.24) and (8.26) in terms of the derivatives of the expansion rate, H , or of the potential, V .

8.2.2.2 Definitions

To formalize the previous argument, let us introduce the slow-roll parameters via the relations

$$\varepsilon = -\frac{\dot{H}}{H^2}, \quad (8.37)$$

$$\delta = \varepsilon - \frac{\dot{\varepsilon}}{2H\varepsilon}, \quad (8.38)$$

$$\xi = \frac{\dot{\varepsilon} - \dot{\delta}}{H}. \quad (8.39)$$

The parameter δ is sometime called η in the literature but we shall not use this notation, to avoid any confusion with the conformal time. The definitions (8.37)–(8.39) have the advantage of depending only on the space-time geometry, i.e. they are not restricted to the single scalar field case. Yet, in practice we shall almost always consider the case of a single scalar field. Using the equations of motion we find that the slow-roll parameters can be rewritten as

$$\varepsilon = \frac{3}{2}\dot{\varphi}^2 \left[\frac{1}{2}\dot{\varphi}^2 + V(\varphi) \right]^{-1}, \quad (8.40)$$

$$\delta = -\frac{\ddot{\varphi}}{H\dot{\varphi}}. \quad (8.41)$$

The Friedmann equations imply that the parameter ε can be expressed as

$$\varepsilon = \frac{1}{4\pi G_N} \left[\frac{H'(\varphi)}{H(\varphi)} \right]^2, \quad (8.42)$$

with the notation $H' \equiv dH/d\varphi$. This indicates that ε is related to the small parameter that appeared in the condition (8.35). It can be used to rewrite the two Friedmann equations as

$$H^2 \left(1 - \frac{1}{3}\varepsilon \right) = \frac{8\pi G_N}{3}V, \quad \frac{\ddot{a}}{a} = H^2 (1 - \varepsilon), \quad (8.43)$$

and the effective equation of state of the inflaton as

$$w_\varphi = -1 + \frac{2}{3}\varepsilon. \quad (8.44)$$

The condition for inflation therefore reduces to

$$\ddot{a} > 0 \iff w < -1/3 \iff \varepsilon < 1. \quad (8.45)$$

The three parameters (8.37)–(8.39) are related via the equations of motion as

$$\frac{\dot{\varepsilon}}{H} = 2\varepsilon(\varepsilon - \delta), \quad \frac{\dot{\delta}}{H} = 2\varepsilon(\varepsilon - \delta) - \xi. \quad (8.46)$$

The slow-roll conditions are thus satisfied as long as ε and δ are small compared to 1 and $\xi = \mathcal{O}(\varepsilon^2, \delta^2, \varepsilon\delta)$. The parameter δ also takes the form

$$\delta = \frac{1}{4\pi G_N} \frac{H_{,\varphi\varphi}}{H}. \quad (8.47)$$

It is also useful to express these parameters in terms of conformal time as

$$\varepsilon = 1 - \frac{\mathcal{H}'}{\mathcal{H}^2}, \quad \delta = 1 - \frac{\varphi''}{\mathcal{H}\varphi'} = \varepsilon - \frac{\varepsilon'}{2\mathcal{H}\varepsilon}. \quad (8.48)$$

8.2.2.3 Expansion of the potential and of the expansion rate

The slow-roll approximation can be seen as a Taylor expansion of the potential V or of the expansion rate H in terms of their derivatives with respect to φ . Two sets of slow-roll parameters can be introduced, based either on the potential V and its derivatives,

$$\varepsilon_V = \frac{1}{16\pi G_N} \left(\frac{V_{,\varphi}}{V} \right)^2, \quad \eta_V = \frac{1}{8\pi G_N} \left(\frac{V_{,\varphi\varphi}}{V} \right), \quad \xi_V = \frac{1}{8\pi G_N} \sqrt{\frac{V_{,\varphi} V_{,\varphi\varphi\varphi}}{V^2}}, \quad (8.49)$$

or on the expansion rate H ,

$$\varepsilon_H = \frac{1}{4\pi G_N} \left(\frac{H_{,\varphi}}{H} \right)^2, \quad \delta_H = \frac{1}{4\pi G_N} \left(\frac{H_{,\varphi\varphi}}{H^2} \right), \quad \xi_H = \frac{1}{4\pi G_N} \sqrt{\frac{H_{,\varphi} H_{,\varphi\varphi\varphi}}{H^2}}. \quad (8.50)$$

Both these sets of parameters appear in the literature. They are not independent and one can show, using the Jacobi equation (8.32), that

$$\varepsilon = \varepsilon_V = \varepsilon_H, \quad \delta = \delta_H = \eta_V - \varepsilon_V, \quad \xi^2 = \xi_H^2 - 3\varepsilon\delta + 2\varepsilon^2, \quad \xi_H^2 = \xi_V^2 - 3\varepsilon\eta_V + 3\varepsilon^2.$$

(8.51)

8.2.2.4 Duration of inflation and number of e-folds

The expansion between an arbitrary time t and the end of inflation can be related to

$$N(t, t_f) = \int_t^{t_f} H dt, \quad (8.52)$$

since, after integration,

$$a(t) = a_f e^{-N}, \quad (8.53)$$

where N , the number of *e-folds*, characterizes the amount of inflation between a given time and the end of the phase of accelerated expansion [$N(t_i, t_f)$ corresponds to the quantity (8.5)]. One can express N in terms of the inflaton as

$$N(\varphi) \equiv N(\varphi, \varphi_f) = \int_\varphi^{\varphi_f} \frac{H}{\dot{\varphi}} d\varphi = -4\pi G_N \int_\varphi^{\varphi_f} \frac{H}{H'} d\varphi = -\sqrt{4\pi G_N} \int_\varphi^{\varphi_f} \frac{d\varphi}{\sqrt{\varepsilon}}, \quad (8.54)$$

where the explicit reference to the end of inflation φ_f in N is dropped when there is no ambiguity.

During inflation, a mode k becomes super-Hubble at the time t_k defined by

$$k = a(t_k)H(t_k).$$

We will note with an index k , for example, X_k , any quantity X evaluated at t_k , $X(t_k) = X_k$. The dynamics of inflation can thus be described using either t , φ or k as the evolution variable. For instance, we will denote by $N(\varphi)$ and N_k the number of e-folds between the moment where the inflaton takes the value φ or that where the scale k becomes super-Hubble and the end of inflation.

It is therefore useful to relate k to the value of the inflaton at t_k by

$$k(\varphi) = a_f H(\varphi) \exp[-N(\varphi)], \quad (8.55)$$

so that the speed at which scales become super-Hubble is characterized by

$$\frac{d \ln k}{d \varphi} = \sqrt{\frac{4\pi G_N}{\varepsilon}} (\varepsilon - 1). \quad (8.56)$$

Similarly, one can easily show that the relation $k = a_k H_k$ implies

$$\frac{d \ln H_k}{d \ln k} = -\varepsilon. \quad (8.57)$$

The first expression points out that if $H' = 0$ (and thus $\varepsilon = 0$), i.e. for a pure de Sitter phase, $d \ln k / d \varphi = 0$ and the process of Hubble radius crossing is scale invariant.

8.2.2.5 Sensitivity to initial conditions

The Klein–Gordon equation, being a second-order differential equation, has two independent solutions. In the slow-roll regime, however, it becomes a first-order differential equation, with only one solution. One of the solutions has thus been neglected. In the Hamilton–Jacobi formalism, the solution of (8.32) depends on the choice of the initial condition for the scalar field. In order for the results we obtain to be robust, and for inflation to be predictive, the late times solution $H(\varphi)$ should be independent of the choice of initial conditions [14].

To see this, let us assume that $H_0(\varphi)$ is the slow-roll solution and let us consider a small perturbation with respect to this solution, $H(\varphi) = H_0 + \delta H$. By linearizing (8.32), we find that the perturbation δH is a solution of $H'_0 \delta H' \simeq 3H_0 \delta H / 12\pi G_N$ and takes the general form

$$\delta H = \delta H(\varphi_i) \exp \left[12\pi G_N \int_{\varphi_i}^{\varphi} \frac{H_0(\varphi)}{H'_0(\varphi)} d\varphi \right],$$

where $\delta H(\varphi_i)$ represents a different choice of initial condition. The previous integral can be expressed in terms of the number of e-folds (8.54) as

$$\delta H = \delta H(\varphi_i) \exp \{-3[N(\varphi_i) - N(\varphi)]\}. \quad (8.58)$$

The effect of the choice of initial conditions is therefore exponentially erased and has no observable effect as soon as a few e-folds of inflation have occurred before the scales of cosmological interest become super-Hubble.

Another way to illustrate that the slow-rolling solution is an attractor of the dynamics is to depict it as a phase portrait in the $(\varphi, \dot{\varphi})$ -plane, starting from arbitrary initial conditions. Fig. 8.4 gives an example of such a phase portrait for a potential $V = m^2\varphi^2/2$. The spiralling around the origin describes the dynamics at the end of the inflationary period and is relevant for the reheating that will be discussed later.

We also see in the picture that when the inflationary period is short, there is a possibility that the actual solution has not converged toward the inflationary attractor at the time the modes of observational interest are becoming super-Hubble. In such a case the predictions of inflation will be discussed in terms of trajectories.

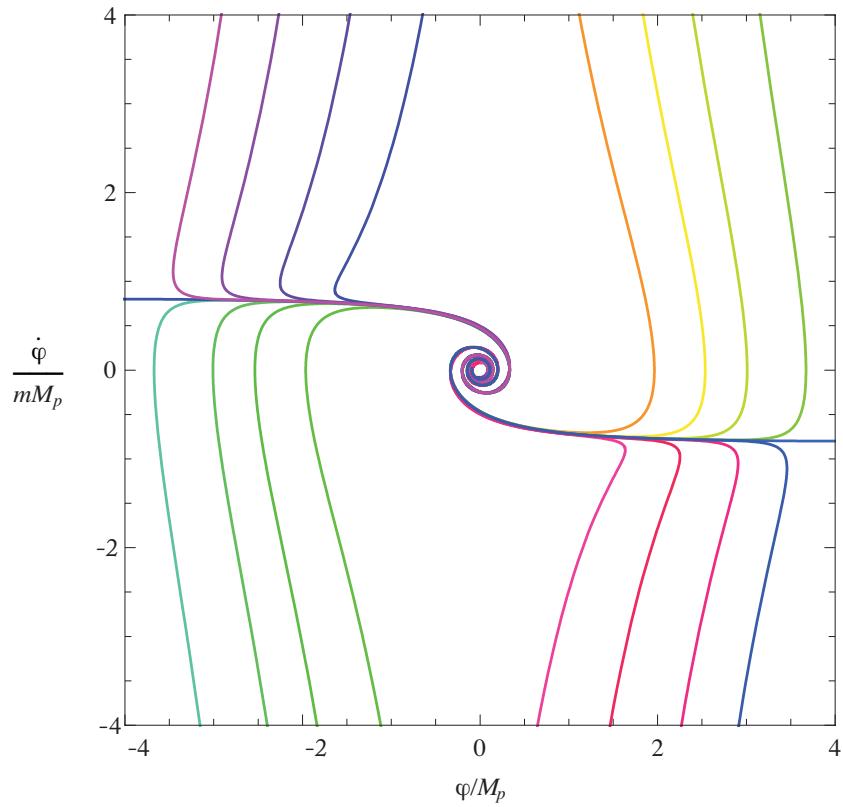


Fig. 8.4 Phase portrait in the $(\varphi, \dot{\varphi})$ -plane of the dynamics of a scalar field with a potential $V = m^2\varphi^2/2$, assuming $m = 10^{-6}M_p$. This illustrates the mechanism of attraction toward the slow-roll solution.

8.2.3 End of inflation

The end of inflation can be identified with the end of the slow-roll regime. We can therefore define the value of the field at the end of inflation, φ_f , by

$$\max(\varepsilon, \delta) \simeq 1. \quad (8.59)$$

At the end of inflation, all classical inhomogeneities have been exponentially washed out, and one can consider them as non-existent at this stage. The Universe has become very flat so that curvature terms can be neglected and one can set $K = 0$ during all the primordial phase (see Section 8.9 for a discussion of the effect of spatial curvature). Moreover, all the entropy has been diluted. If the inflaton potential has a minimum, the scalar field will oscillate around this minimum right after the end of inflation (see Fig. 8.4).

Due to the Hubble expansion, these oscillations are damped and the scalar field decays into a large number of particles. During this phase, the inflationary Universe (of low entropy and dominated by the coherent oscillations of the inflaton) becomes a hot Universe (of high entropy and dominated by radiation). This *reheating* phase, described in Section 8.6, connects inflation with the hot Big-Bang scenario.

To analyse the transition from the inflating Universe to the observable Universe, it will be useful to relate the instant at which a mode k becomes super-Hubble during inflation to that where it becomes sub-Hubble again. In order to do so, one needs a complete model of the evolution of the Universe. In the simplest case, the sequence of events is: inflation – matter era during the reheating phase (denoted with the index ‘reh’) – radiation era – matter era (see Fig. 8.5). Assuming that the transitions between these various regimes are instantaneous, one has

$$\begin{aligned} \frac{k}{a_0 H_0} &= \frac{a_k H_k}{a_0 H_0} \\ &= \frac{a_k}{a_f} \frac{a_f}{a_{\text{reh}}} \frac{a_{\text{reh}}}{a_{\text{eq}}} \frac{a_{\text{eq}}}{a_0} \frac{H_k}{H_0}. \end{aligned} \quad (8.60)$$

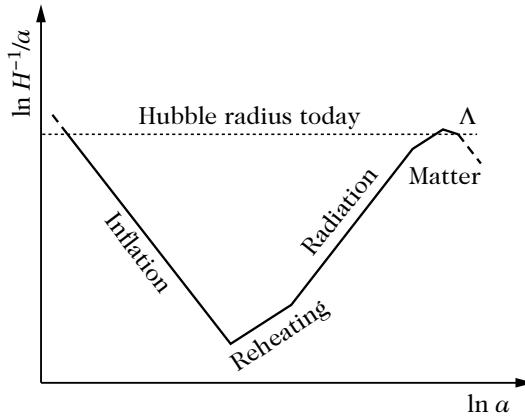


Fig. 8.5 The function $\ln aH$ as a function of $\ln a$ illustrates the different epochs to take into account to compute the number of e-folds.

Using the typical values of these epochs, we obtain

$$N(k) \simeq 62 - \ln \left(\frac{k}{a_0 H_0} \right) - \ln \left(\frac{10^{16} \text{ GeV}}{V_k^{1/4}} \right) + \frac{1}{4} \ln \left(\frac{V_k}{V_f} \right) - \frac{1}{3} \ln \left(\frac{V_f^{1/4}}{\rho_{\text{reh}}^{1/4}} \right) - \ln h. \quad (8.61)$$

This relation can be modified if inflation occurs at a lower-energy scale. Moreover, the mode of wavelength equal to the Hubble radius today, $k_0 = a_0 H_0$, becomes super-Hubble at around 60 e-folds before the end of inflation. A scale of the order of 1 Mpc, typically of the order of the smallest observable scales today, becomes super-Hubble at around 9 e-folds after k_0 . Our observations are thus mainly sensitive to the dynamics of inflation during about ten e-folds.

Finally, let us note that $\Omega_K(k) = -K/a_k^2 H_k^2 = -K/a_0^2 H_0^2 = \Omega_{K0}$ for the largest mode observable today, i.e. k_0 . Constraints obtained in Chapter 4 show that it is completely legitimate to neglect the spatial curvature during inflation for the epochs during which the modes observable today become super-Hubble.

8.2.4 Some examples

8.2.4.1 Chaotic inflation

The simplest model

Chaotic inflation is based on considering a free massive scalar field so that

$$V(\varphi) = \frac{1}{2}m^2\varphi^2. \quad (8.62)$$

The Klein–Gordon equation (8.26) reduces to that of a damped harmonic oscillator, $\ddot{\varphi} + 3H\dot{\varphi} + m^2\varphi = 0$. If φ is initially large, then the Friedmann equation (8.24) implies that H is also very large. The friction term becomes important and dominates the dynamics so that the field must be in the slow-roll regime. The evolution equations then reduce to the following two first-order equations,

$$3H\dot{\varphi} + m^2\varphi = 0, \quad H^2 = \frac{4\pi}{3}\left(\frac{m}{M_p}\right)^2\varphi^2. \quad (8.63)$$

They can be integrated easily to give

$$\varphi(t) = \varphi_i - \frac{mM_p}{\sqrt{12\pi}}t, \quad (8.64)$$

$$a(t) = a_i \exp\left\{\frac{2\pi}{M_p^2} [\varphi_i^2 - \varphi^2(t)]\right\}, \quad (8.65)$$

where φ_i and a_i are the values of the field and the scale factor at $t_i = 0$. The slow-roll parameters (8.37)–(8.39) being given by

$$\varepsilon = \frac{M_p^2}{4\pi\varphi^2}, \quad \delta = 0, \quad \xi_V = 0, \quad (8.66)$$

the slow-roll regime lasts until the field φ reaches $\varphi_f = M_p/\sqrt{4\pi}$. We infer that the total number of e-folds is

$$N(\varphi_i) = 2\pi\left(\frac{\varphi_i}{M_p}\right)^2 - \frac{1}{2}. \quad (8.67)$$

In order to have $N \gtrsim 70$, we need $\varphi_i \gtrsim 3M_p$. If φ takes the largest possible value compatible with the classical description adopted here, i.e. fixed by $V(\varphi_i) \lesssim M_p^4$, we

find that $\varphi_i \sim M_p^2/m$. In this case, the maximal accessible number of e-folds would be $N_{\max} \sim 2\pi M_p^2/m^2$. As will be seen later, the theory of cosmological perturbations and observations of the cosmic microwave background impose that the mass is of the order of $m \sim 10^{-6} M_p$, so that $N_{\max} \sim 10^{13}$. The maximal number of e-folds is thus very large compared to the minimum required for solving the cosmological problems.

Consequently, if the Universe is initially composed of regions where the values of the scalar field are randomly distributed, or if we consider different Universes with different field values, then the domains where the initial value of φ is too small, never inflate or only for a small number of e-folds. The main contribution to the total physical volume of the Universe at the end of inflation thus comes from regions that have inflated for a long time and that had an initially large value of φ . These domains produce extremely flat and homogeneous zones at the end of inflation with a very large size compared to that of the observable Universe. At larger scales, the Universe has a very inhomogeneous structure. We will come back to this chaotic initial condition idea in Section 8.7.

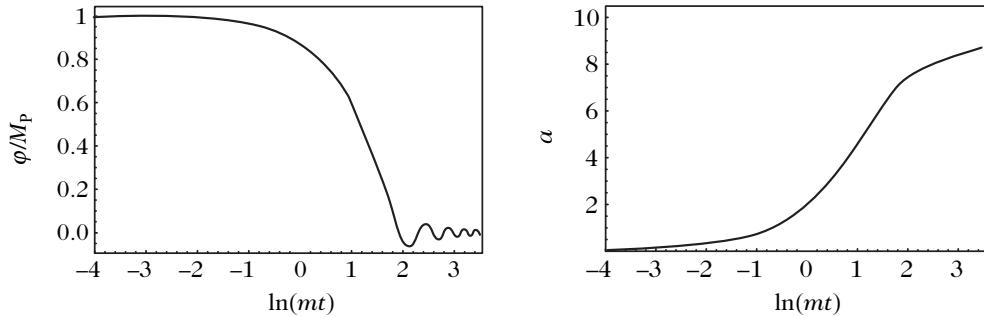


Fig. 8.6 Evolution of the inflaton and scale factor during chaotic inflation for the model (8.62). The scalar field is initially in a slow-roll regime and the expansion of the Universe is accelerated. At the end of this regime, it starts oscillating at the minimum of its potential and it is equivalent to a pressureless fluid.

Extensions

This model can be generalized to any polynomial function of the scalar field. As an example, let us consider the family of potentials

$$V(\varphi) = \frac{3M_p^4}{8\pi} \lambda_n \left(\frac{\varphi}{M_p} \right)^n, \quad (8.68)$$

where λ_n is a dimensionless constant. The slow-roll parameters are given by

$$\varepsilon = \frac{M_p^2}{16\pi} \frac{n^2}{\varphi^2}, \quad \delta = \frac{M_p^2}{16\pi} \frac{n(n-2)}{\varphi^2}, \quad (8.69)$$

and the slow-roll regime lasts as long as $\varphi > \varphi_f = nM_p/4\sqrt{\pi}$. We infer that the total number of e-folds is

$$N(\varphi_i) = \frac{4\pi}{n} \left(\frac{\varphi_i}{M_p} \right)^2 - \frac{n}{4}. \quad (8.70)$$

So in order for $N > N_{\min}$, we should initially have

$$\varphi_i > M_p \sqrt{\frac{n}{4\pi} \left(N_{\min} + \frac{n}{4} \right)}. \quad (8.71)$$

Yet, chaotic initial conditions imply that $V(\varphi_i) \sim M_p^4$, imposing

$$\varphi_i \sim M_p \left(\frac{8\pi}{3\lambda_n} \right)^{1/n}, \quad (8.72)$$

a value in general very large compared to M_p . Solving the Klein–Gordon and Friedmann equations in the slow-roll regime approximation gives

$$\varphi(t) = \varphi_i \left[1 + n(n-4) \frac{\sqrt{\lambda_n}}{16\pi} \left(\frac{\varphi_i}{M_p} \right)^{(n-4)/2} M_p t \right]^{2/(4-n)}, \quad (8.73)$$

$$a(t) = a_i \exp \left\{ \frac{4\pi}{nM_p^2} [\varphi_i^2 - \varphi^2(t)] \right\}. \quad (8.74)$$

The expression in square brackets for the solution of the scalar field is positive definite as long as the slow-roll hypothesis is valid.

Notice that although expression (8.74) is well defined for all n , the scalar field on the other hand is singular for $n = 4$. The direct computation of this specific case shows that (8.74) remains valid, with $\varphi(t) = \varphi_i \exp(-\frac{1}{2\pi} \sqrt{\lambda_4} M_p t)$.

We can reexpress these quantities in terms of the number of e-folds to obtain

$$\varphi(N) = M_p \sqrt{\left(\frac{\varphi_i}{M_p} \right)^2 - \frac{n}{4\pi} N}, \quad H(N) = M_p \sqrt{\lambda_n} \left[\left(\frac{\varphi_i}{M_p} \right)^2 - \frac{n}{4\pi} N \right]^{n/4}, \quad (8.75)$$

valid until $N(\varphi_i)$. These expressions show that the Hubble parameter can vary significantly during the slow-roll phase. In particular, and as will be seen later, the amplitude of the observable modes generated roughly 50–70 e-folds before the end of inflation implies that $H/M_p \sim 10^{-5}$. However, this does not imply that the Hubble parameter could not have been much larger before that. Cosmological observations can thus give some inputs on the last 70 e-folds or so of the inflation era but, a priori, put no constraints on the primordial phase of inflation.

8.2.4.2 Power-law inflation: an exact solution

For most models, the slow-roll approximation is so good that there is no need to go further, even if the end of inflation is not very well taken into account. It is, however, useful to have some exact solutions.

Many models exist in the literature but the most attractive one is that of power-law inflation [15]. The advantage of this model is that the perturbation equations can also be solved analytically.

These models are constructed such that the solution of the equations of evolution are

$$a = a_0 |\eta|^{1+\beta} \iff a = a_0 t^p, \quad 1 + \beta = \frac{p}{1-p}, \quad (8.76)$$

and

$$\frac{\varphi - \varphi_i}{M_p} = \frac{1}{2} \sqrt{\frac{\gamma}{\pi}} (1 + \beta) \ln |\eta|, \quad \gamma = \frac{\beta + 2}{\beta + 1} = \frac{1}{p}. \quad (8.77)$$

The parameter p varies between 1 and ∞ and thus $\beta < -2$. The evolution equations imply that the potential is of the form

$$V(\varphi) = V_i \exp \left[4 \sqrt{\frac{\pi}{p}} \frac{(\varphi - \varphi_i)}{M_p} \right], \quad (8.78)$$

so that the slow-roll parameters turn out to be constant,

$$\varepsilon = \delta = \xi_H = \frac{1}{p}, \quad \xi = 0. \quad (8.79)$$

Thus, with such a potential, the slow-roll parameters' inflation never ends. The limit $p \rightarrow +\infty$ (i.e. $\beta \rightarrow -2$) corresponds to a de Sitter Universe.

8.2.4.3 Hybrid inflation

Models of hybrid inflation introduce two coupled scalar fields, φ and σ , with the potential

$$V(\varphi, \sigma) = \frac{\lambda}{4} (\sigma^2 - v^2)^2 + \frac{1}{2} g^2 \varphi^2 \sigma^2 + V(\varphi), \quad (8.80)$$

where v is the vacuum expectation value (the VEV, as defined in Chapter 2) of σ and $V(\varphi)$ is the potential of the field φ (see Fig. 8.7). The effective mass of the field σ depends on the value of the second field

$$m^2(\sigma) = g^2 (\varphi^2 - \varphi_c^2), \quad \varphi_c^2 = \frac{\lambda v^2}{g^2}.$$

The point $\varphi = \varphi_c$ is the bifurcation point where the effective mass of the field σ changes sign to become negative. As long as $\varphi > \varphi_c$, the potential has a valley along $\sigma = 0$; everything therefore goes as if there was only one field rolling towards the minimum of the potential $V(\varphi) + \lambda v^4/4$. The global minimum is, however, in $\varphi = 0$, $|\sigma| = v$. When the inflaton reaches the value φ_c , the mass of σ changes sign, which results in a spinodal instability (see Chapter 11) triggering the end of inflation. The total number of e-folds is then

$$N = 2\pi \frac{\lambda v^4}{m^2 M_p^2} \ln \frac{\varphi_i}{\varphi_c}, \quad (8.81)$$

if the potential for the inflaton is of the form $V(\varphi) = m^2 \varphi^2/2$.

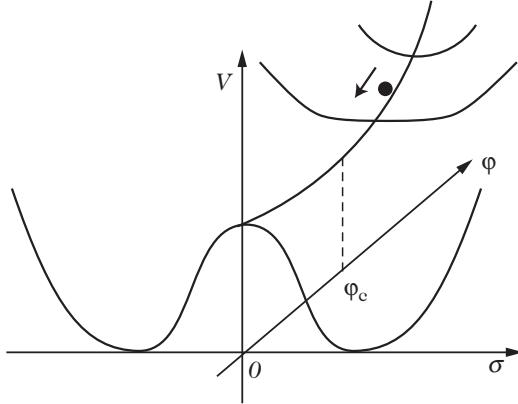


Fig. 8.7 Potential for hybrid inflation with two scalar fields. Inflation occurs when $\sigma = 0$ and stops abruptly when φ reaches the value φ_c of the bifurcation point.

8.2.5 Classification of inflationary models

For single-field inflationary models, the potential can always be characterized by two mass scales, Λ and μ ,

$$V(\varphi) = \Lambda^4 f\left(\frac{\varphi}{\mu}\right). \quad (8.82)$$

These two scales typically determine the height and width of the potential.

In order to discuss the phenomenology of these models, it is useful to classify them as a function of the value of their slow-roll parameters. We distinguish between the four following classes.

- A. *Inflation with large values of the inflaton:* $-\varepsilon < \delta \leq \varepsilon$. The archetypes of this class are the potentials of chaotic inflation, e.g. (8.62) and (8.68) or exponential potentials (8.78). In these models, the field is initially away from its minimum towards which it slowly relaxes. These potentials satisfy $V'' > 0$, i.e. $\eta_V > 0$. In the framework of chaotic inflation, the condition $V \sim M_p^4$ implies that $\varphi \gg M_p$, but a value $\varphi \sim M_p$ is, in general, sufficient to obtain the required number of e-folds.
- B. *Inflation with small values of the inflaton:* $\delta < -\varepsilon$. These potentials are those that appear naturally in the mechanisms of spontaneous symmetry breaking and were at the origin of the so-called models of new inflation. The field is initially close to an unstable equilibrium position in $\varphi_i = 0$. Inflation occurs when the field is close to the origin and the asymptotic form of the potential is not significant. We therefore have $V'' < 0$, i.e. $\eta_V < 0$ and $\varepsilon > 0$ and very small (very flat potential). We infer that $\delta < -\varepsilon$. An example of such a potential is

$$V(\varphi) = \Lambda^4 \left[1 - \left(\frac{\varphi}{\mu} \right)^p \right], \quad (8.83)$$

that we can consider as a Taylor expansion of the ‘true’ potential around the origin. In the case of a spontaneous symmetry-breaking potential of the Higgs

kind, $p = 2$; μ is then the scale of the symmetry breaking. Another example is the Coleman–Weinberg potential,

$$V = \frac{\lambda}{4}\varphi^4 + \frac{\lambda^2}{44\pi}\varphi^4 \left[\ln\left(\frac{\varphi}{\mu}\right) - \frac{25}{6} \right]. \quad (8.84)$$

- C. *Hybrid models*: $0 < \delta < \varepsilon$. These potentials often appear in the framework of supersymmetry. The field evolves towards its non-vanishing minimum [cf. the example given in (8.80)]. Inflation usually ends abruptly due to an instability of the auxiliary field. Another example is given by the potential $V(\varphi) = \Lambda^4[1 + (\varphi/\mu)^p]$.
- D. *Linear models*: $\delta = -\varepsilon$. These models, characterized by

$$V(\varphi) \propto \varphi, \quad (8.85)$$

represent a limit between the models with large and small values of the field.

Of course, such a classification is not exhaustive as it does not incorporate potentials of the type $V(\varphi) \propto \ln \varphi$, which often appear in the framework of supersymmetry, or of inverse power laws, $V(\varphi) \propto \varphi^{-n}$ as in intermediate inflation. However, these models require the introduction of auxiliary fields to stop inflation.

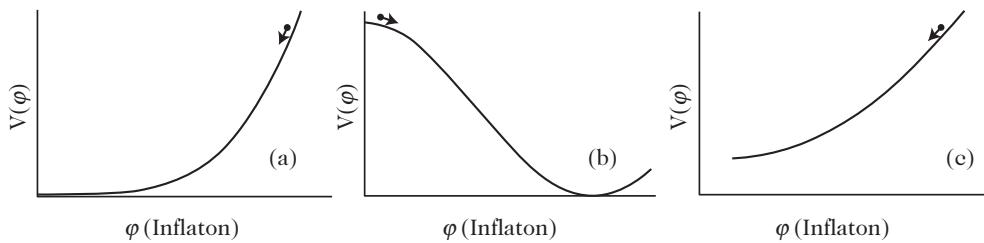


Fig. 8.8 From left to right, three examples of inflationary potentials with large field values (class A), small field values (class B) and hybrid (class C).

8.3 Quantum fluctuations during inflation

One of the great successes of inflation has been to provide a way to relate the origin of the large-scale structure of the Universe to the quantum fluctuations of the inflaton that are amplified during inflation. To address this central point of primordial cosmology, we start by a warm up and only consider test fields.

This study will illustrate the fact that during inflation, any light field (in the sense $m < H$) develops super-Hubble fluctuations with mean amplitude $H/2\pi$.

8.3.1 Massless test scalar field in a de Sitter space-time

The structure of the vacuum in an exponentially expanding space-time is much more complex than that in a Minkowski space-time (Chapter 2). In particular, the wavelength of any fluctuation grows exponentially. When it becomes larger than the Hubble

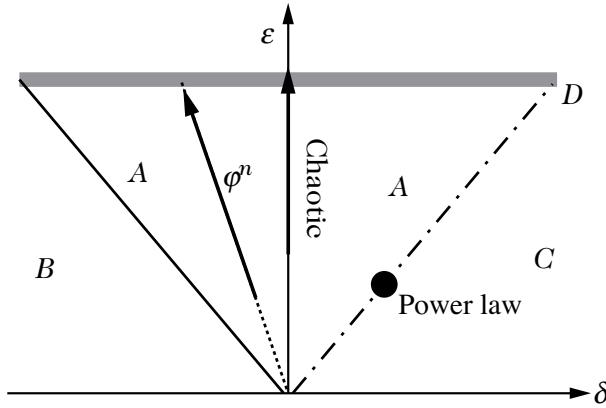


Fig. 8.9 The four classes of models of single-field inflation in the space of the two first slow-roll parameters. We have indicated the trajectory of the models of chaotic inflation (8.62) and (8.68) and the point represents a model of power-law inflation (8.78).

radius, H^{-1} , this fluctuation is said to freeze as it no longer oscillates due to the friction term in the Klein–Gordon equation, and it rapidly converges toward a solution of constant amplitude.

8.3.1.1 Quantization

To quantize a scalar field in an expanding space-time, we proceed in an analogous way to what is usually done in Minkowski space-time and promote the scalar field, χ say, to the status of a quantum operator,

$$\hat{\chi}(\mathbf{x}, t) = \int \frac{d^3\mathbf{k}}{(2\pi)^{3/2}} \left[\chi_k(t) e^{i\mathbf{k}\cdot\mathbf{x}} \hat{a}_{\mathbf{k}} + \chi_k^*(t) e^{-i\mathbf{k}\cdot\mathbf{x}} \hat{a}_{\mathbf{k}}^\dagger \right], \quad (8.86)$$

where the creation and annihilation operators satisfy (2.83). The mode function χ_k is a solution of the classical Klein–Gordon equation (1.108)

$$\ddot{\chi}_k + 3H\dot{\chi}_k + \frac{k^2}{a^2}\chi_k = 0. \quad (8.87)$$

Qualitatively, this equation possesses two regimes. When the mode is sub-Hubble, $k \gg aH$, the friction term is negligible and the equation reduces to that of a harmonic oscillator analogous to that obtained in Minkowski space-time, with the only difference being that the wave number is redshifted by the expansion. When the mode is super-Hubble, $k \ll aH$, and the gradient term is negligible. The equation reduces to $\ddot{\chi}_k + 3H\dot{\chi}_k = 0$ where the dominant solution is a constant. In conclusion, an initially sub-Hubble mode oscillates as its wavelength grows and then freezes when it becomes super-Hubble.

8.3.1.2 General solution

To solve (8.87) it is convenient to work in conformal time and to introduce the auxiliary field

$$v = a\chi. \quad (8.88)$$

The Klein–Gordon equation then takes the form

$$v_k'' + \left(k^2 - \frac{a''}{a} \right) v_k = 0,$$

(8.89)

which introduces an effective time-dependent mass term, $m^2 = -a''/a$.

For a de Sitter space, the scale factor is given by (8.14) so that v satisfies

$$v_k'' + \left(k^2 - \frac{2}{\eta^2} \right) v_k = 0. \quad (8.90)$$

It has the general solution

$$v_k(\eta) = \left[A(k)H_{3/2}^{(1)}(-k\eta) + B(k)H_{3/2}^{(2)}(-k\eta) \right] \sqrt{-\eta}, \quad (8.91)$$

where $H_{3/2}^{(1,2)}$ are the Hankel functions of first and second kind (Appendix B). Using

$$H_{3/2}^{(2)}(z) = \left[H_{3/2}^{(1)}(z) \right]^* = -\sqrt{\frac{2}{\pi z}} e^{-iz} \left(1 + \frac{1}{iz} \right), \quad (8.92)$$

we infer the general form

$$v_k(\eta) = A(k)e^{-ik\eta} \left(1 + \frac{1}{ik\eta} \right) + B(k)e^{ik\eta} \left(1 - \frac{1}{ik\eta} \right). \quad (8.93)$$

At this stage, we need to postulate a choice of initial conditions to determine the two arbitrary functions $A(k)$ and $B(k)$.

8.3.1.3 Initial condition

To perform a canonical quantization, we impose the commutation rules on the operators \hat{v} and $\hat{\pi}$, namely $[\hat{v}(\mathbf{x}, \eta), \hat{v}(\mathbf{x}', \eta)] = [\hat{\pi}(\mathbf{x}, \eta), \hat{\pi}(\mathbf{x}', \eta)] = 0$ and $[\hat{v}(\mathbf{x}, \eta), \hat{\pi}(\mathbf{x}', \eta)] = \delta^{(3)}(\mathbf{x} - \mathbf{x}')$ on constant time hypersurfaces, $\hat{\pi}$ being the conjugate momentum of \hat{v} . From the commutation rules of the annihilation and creation operators, this implies that

$$v_k v_k'^* - v_k^* v_k' = i, \quad (8.94)$$

which determines the normalization of the Wronskian. The choice of a specific mode function $v_k(\eta)$ corresponds to the choice of a prescription for the physical vacuum $|0\rangle$, defined by

$$\hat{a}_{\mathbf{k}}|0\rangle = 0.$$

To choose the vacuum, let us note that for very high frequency modes, $k\eta \gg 1$, the curvature term is negligible so that the field can be quantized as if it were in a

Minkowski space-time (see Chapter 2). We thus pick up the solution that corresponds adiabatically to the usual Minkowski vacuum, keeping only the positive frequencies. We therefore impose

$$v_k \rightarrow \frac{e^{-ik\eta}}{\sqrt{2k}}, \quad (8.95)$$

when $k\eta \rightarrow -\infty$. This choice is referred to as the Bunch–Davies vacuum.² We infer that the solution satisfying this initial condition is

$$\boxed{\chi_k = \frac{H\eta}{\sqrt{2k}} \left(1 + \frac{1}{ik\eta} \right) e^{-ik\eta}.} \quad (8.96)$$

Let us emphasize that, at this stage, we have not justified why we have to apply the condition (8.95) to v_k rather than to χ_k . Such a justification will be made later.

8.3.1.4 Properties on super-Hubble scales

On super-Hubble scales, the field χ acquires a constant amplitude

$$|\chi_k| (k\eta \ll 1) = \frac{H}{\sqrt{2k^3}}, \quad (8.97)$$

where we have used $a(\eta) = -1/(H\eta)$ for a pure de Sitter inflationary phase, H being a constant in this case.

Let us emphasize that, in this limit, the field actually behaves as

$$\hat{\chi} \rightarrow \int \frac{d^3 k}{(2\pi)^{3/2}} \hat{\chi}_k e^{i\mathbf{k}\cdot\mathbf{x}} = \int \frac{d^3 k}{(2\pi)^{3/2}} \frac{H}{\sqrt{2k^3}} (\hat{a}_k + \hat{a}_{-k}^\dagger) e^{i\mathbf{k}\cdot\mathbf{x}}.$$

At this point, all the modes being proportional to $(\hat{a}_k + \hat{a}_{-k}^\dagger)$, the quantum operators $\hat{\chi}_k$ commute. Therefore, $\hat{\chi}$ has actually the same statistical properties as a Gaussian classical stochastic field. Effectively, on super-Hubble scales, the quantum operator $\hat{\chi}$ can be replaced by a stochastic field with Gaussian statistics. Introducing a unit Gaussian random variable, $e_v(\mathbf{k})$, which satisfies

$$\langle e_v(\mathbf{k}) \rangle = 0, \quad \langle e_v(\mathbf{k}) e_v^*(\mathbf{k}') \rangle = \delta^{(3)}(\mathbf{k} - \mathbf{k}'),$$

the mode operators are replaced by stochastic variables through

$$\hat{v}_k \rightarrow v_k = v_k(\eta) e_v(\mathbf{k})$$

and we can switch from the vacuum quantum average $\langle 0 | \dots | 0 \rangle$ to the classical ensemble average $\langle \dots \rangle$.

This explains why the super-Hubble modes have Gaussian statistics. A description of the quantum to classical transition can be found in Ref. [16].

²Note that the quantum states have a clear physical interpretation only once v_k has been chosen. The normalization condition (8.94) does not fix v_k completely and any function of the form $u_k = \alpha_k v_k + \beta_k v_k^*$, with α_k and β_k two complex numbers satisfying $|\alpha_k|^2 - |\beta_k|^2 = 1$ will also be a solution normalized according to (8.94). The asymptotic condition (8.95) completely defines the quantum states.

8.3.1.5 Power spectrum

The correlation function of v is defined as

$$\xi_v \equiv \langle 0 | \hat{v}(\mathbf{x}, \eta) \hat{v}(\mathbf{x}', \eta) | 0 \rangle,$$

and takes the simple form

$$\xi_v = \int \frac{d^3 k}{(2\pi)^3} |v_k|^2 e^{i\mathbf{k} \cdot (\mathbf{x} - \mathbf{x}')} = \int \frac{dk}{k} \frac{k^3}{2\pi^2} |v_k|^2 \frac{\sin kr}{kr}, \quad (8.98)$$

where the isotropy of the background space-time has been used to integrate over angles. Because of the symmetries of the background, it is a function of $r = |\mathbf{x} - \mathbf{x}'|$ only. We thus define the power spectra of v , using the definitions of Section B.5 in Appendix B for the normalization of the power spectrum,

$$P_v(k) = |v_k|^2, \quad \mathcal{P}_v(k) = \frac{k^3}{2\pi^2} |v_k|^2. \quad (8.99)$$

In the stochastic picture, the correlator of $v_{\mathbf{k}}$ is simply given by

$$\langle v_{\mathbf{k}} v_{\mathbf{k}'}^* \rangle = P_v(k) \delta^{(3)}(\mathbf{k} - \mathbf{k}'), \quad (8.100)$$

from which one easily deduces that the power spectrum of χ is

$$P_{\chi}(k) = \frac{2\pi^2}{k^3} \mathcal{P}_{\chi}(k) = |\chi_k|^2 = \frac{|v_k|^2}{a^2}.$$

It is thus a scale invariant power spectrum,

$$\mathcal{P}_{\chi}(k) = \left(\frac{H}{2\pi} \right)^2. \quad (8.101)$$

8.3.1.6 WKB approximation

In more general cases, where e.g. the function $a(\eta)$ is not known analytically, we may not have exact solutions for the mode functions. One can generalize the previous construction by relying on a WKB approach in which one introduces the WKB mode functions

$$v_k^{\text{WKB}}(\eta) = \frac{1}{\sqrt{2\omega(k, \eta)}} e^{\pm i \int^{\eta} \omega(k, \eta') d\eta'}, \quad (8.102)$$

which, as is easily checked, are solutions of

$$v_k^{\text{WKB}''} + [\omega^2(k, \eta) - Q_{\text{WKB}}] v_k^{\text{WKB}} = 0,$$

where the potential is

$$Q_{\text{WKB}} = \frac{3}{4} \left(\frac{\omega'}{\omega} \right)^2 - \frac{1}{2} \frac{\omega''}{\omega}. \quad (8.103)$$

The WKB solution is thus a good approximation as long as the WKB condition

$$\left| \frac{Q_{\text{WKB}}}{\omega^2} \right| \ll 1 \quad (8.104)$$

holds. For a function satisfying an equation such as (8.90),

$$\omega^2(k, \eta) = k^2 - \frac{2}{\eta^2},$$

the WKB condition reduces to $|k\eta| \gg 1$, so that on sub-Hubble scales, the mode function is actually close to its WKB approximation.

One might thus reinterpret the quantization procedure by noting the fact that there exists an adiabatic (WKB) solution on sub-Hubble scales.

8.3.1.7 Interpretation and infra-red divergences

Expression (8.96) can be used to express the variance of the field, $\langle \chi^2 \rangle$, as a function of the physical momentum, $\mathbf{p} = a^{-1}(t)\mathbf{k}$,

$$\langle \chi^2 \rangle = \int \frac{d^3 p}{(2\pi)^3 p} \left(\frac{1}{2} + \frac{H^2}{2p^2} \right). \quad (8.105)$$

The first term is simply the zero-point energy of a harmonic oscillator in Minkowski space-time (Chapter 2). This term leads to an ultraviolet divergence that can be eliminated by renormalization. The second term only appears in a curved space-time and is characteristic of de Sitter space-time. It can be interpreted as the existence of

$$n_p = \frac{H^2}{2p^2} \quad (8.106)$$

particles of physical momentum p . This term diverges in the infra-red. There are actually two divergences, one when $p \rightarrow 0$ and the second one when $\eta \rightarrow 0$.

The infra-red divergence at $p = 0$ appears because the expansion of the field in creation and annihilation modes do not include the zero mode. For a massless field in Minkowski space in a box of volume V , a complete description of the solutions requires the introduction of the position and momentum operators, $\hat{\varphi}_0$ and $\hat{\pi}_{\varphi 0}$, that satisfy the relation $[\hat{\varphi}_0, \hat{\pi}_{\varphi 0}] = i$. We can then define the vacuum by imposing $\hat{\pi}_{\varphi 0}|0\rangle = 0$ and $\hat{a}_{\mathbf{k}}|0\rangle = 0$, which amounts to describing the state in a space that is the product of a Fock space and a Hilbert space. In a Minkowski space of dimension larger than 2, the continuum limit ($V \rightarrow \infty$) exists because the phase space volume factor compensates the divergence. However, it persists in a de Sitter space-time.

If this divergence is regularised by imposing a frequency cutoff $p > p_0 \sim H$ then

$$\langle \chi^2 \rangle \sim \frac{H^2}{4\pi^2} \int_{H e^{-H t}} \frac{dk}{k} \sim \frac{H^3}{4\pi^2 t}. \quad (8.107)$$

This result is general and the resolution of the infra-red divergence at $k = 0$ does not resolve this second problem. The physical origin of this divergence is found in the eternal expansion that brings modes towards $p = 0$ exponentially.

However, these divergences only appear for massless particles in a de Sitter space that characterizes a phase of eternal inflation and naturally disappear in quasi-de Sitter space-times.

8.3.1.8 Main conclusions

We conclude that during a de Sitter inflationary phase, a test scalar field develops super-Hubble fluctuations with a scale invariant power spectrum and Gaussian statistics. They arise from the parametric amplification of the inescapable vacuum fluctuation of the scalar field on sub-Hubble scales. Their amplitude is dictated by H , that is typically by the energy scale, $2H^2M_p^2/8\pi$, during inflation.

As we shall see, most of these conclusions generalize to more sophisticated cases, and in particular to the inflaton.

8.3.2 Massive test field in de Sitter

We can extend the previous case to that of a test field of mass m_χ . Equation (8.89) then takes the form

$$v_k'' + \left(k^2 + \frac{m^2/H^2 - 2}{\eta^2} \right) v_k = 0, \quad (8.108)$$

when considering a potential $m_\chi^2\chi^2/2$. It can be rewritten as

$$v_k'' + \left[k^2 + \frac{1}{\eta^2} \left(\nu_\chi^2 - \frac{1}{4} \right) \right] v_k = 0, \quad \nu_\chi^2 = \frac{9}{4} - \frac{m_\chi^2}{H^2}. \quad (8.109)$$

If ν_χ is real, the general solution is a generalization of (8.91) and is

$$v_k(\eta) = \left[A(k) H_{\nu_\chi}^{(1)}(-k\eta) + B(k) H_{\nu_\chi}^{(2)}(-k\eta) \right] \sqrt{-\eta}, \quad (8.110)$$

where $H_{\nu_\chi}^{(1,2)}$ are Hankel functions of the first and second kind, respectively, (see Appendix B). The asymptotic form of the Hankel functions allows us to deduce that the solution that satisfies the initial condition (8.95) when $k\eta \rightarrow -\infty$ is

$$v_k = \sqrt{\frac{-\pi\eta}{4}} i^{(\nu_\chi+1/2)} H_{\nu_\chi}^{(1)}(-k\eta) = \sqrt{\frac{-\pi\eta}{4}} i^{(\nu_\chi+1/2)} H_{\nu_\chi}^{(2)}(k\eta), \quad \nu_\chi^2 \geq 0. \quad (8.111)$$

On super-Hubble scales, the field freezes to a constant amplitude,

$$|\chi_k|(k\eta \ll 1) = \frac{H}{\sqrt{2k^3}} 2^{\nu_\chi-3/2} \frac{\Gamma(\nu_\chi)}{\Gamma(3/2)} \left(\frac{k}{aH} \right)^{3/2-\nu_\chi}, \quad (8.112)$$

and has a power spectrum

$$\mathcal{P}_\chi(k) = 2^{2\nu_\chi-3} \left[\frac{\Gamma(\nu_\chi)}{\Gamma(3/2)} \right]^2 \left(\frac{H}{2\pi} \right)^2 \left(\frac{k}{aH} \right)^{3-2\nu_\chi}, \quad (8.113)$$

using the definitions from Section B.5 of Appendix B for the normalization of the power spectrum.

If the field has a mass larger than $3H/2$ then $\nu_\chi^2 < 0$. The solution can be obtained by exchanging ν_χ with $i|\nu_\chi|$ in (8.111). The sub-Hubble solution is not affected and the

initial conditions are therefore identical. Nevertheless, the mode function in the super-Hubble limit is also oscillating. After some algebra, we find that the power spectrum is now given by

$$\mathcal{P}_\chi(k) \propto \left(\frac{H}{2\pi}\right)^2 \left(\frac{H}{m_\chi}\right) \left(\frac{k}{aH}\right)^3. \quad (8.114)$$

The amplitude of the spectrum is thus reduced by a factor $H/m_\chi > 3/2$ and the spectrum decreases rapidly for long wavelengths.

This result is one of the generic predictions of inflation, practically independent of the model at hand.

During inflation, any *light* test field, $m < H$, develops super-Hubble fluctuations of characteristic amplitude $H/2\pi$, with an almost scale invariant power spectrum.

8.3.3 Massive test field during slow-roll inflation

In the two previous examples, inflation was described as a pure de Sitter phase so that H was constant. Of course, this is only an approximation, and in a more realistic model, H decreases slowly in time during inflation.

This time dependence will have an influence on the spectrum of any light field. For a massless scalar field, expression (8.101) means that the spectrum of the super-Hubble modes is given by the value of H at the time where the mode becomes super-Hubble, $\mathcal{P}_\chi(k) \propto H_k^2$. In a de Sitter space-time, H is constant so that H_k does not depend on k and thus the power spectrum is scale invariant. Relation (8.57) indicates that in the slow-roll regime, $H_k \propto k^{-\varepsilon}$, so that we expect that the spectrum of a massless test scalar field behaves as $\mathcal{P}_\chi \propto k^{-2\varepsilon}$. Its spectrum will therefore be *red-tilted* (since $\varepsilon \geq 0$) if inflation is realized in the slow-roll regime.

8.3.3.1 Evolution of the scale factor

The expansion of the Universe is now quasi-de Sitter. Integrating by parts, the conformal time is given by

$$\eta = \int \frac{da}{a^2 H} = -\frac{1}{aH} + \int \frac{\varepsilon}{a^2 H} da. \quad (8.115)$$

Assuming ε is constant, we find that the scale factor evolves as

$$a(\eta) = -\frac{1}{H\eta} \frac{1}{1-\varepsilon}. \quad (8.116)$$

From (8.46), this hypothesis amounts to neglecting all terms of 2nd order in the slow-roll parameters since $\dot{\varepsilon} = \mathcal{O}(\varepsilon^2, \delta\varepsilon)$. This *first-order* slow-roll approximation thus amounts to considering only the two parameters ε and δ and to assume that they are constant.

8.3.3.2 Solution of the Klein–Gordon equation

To solve (8.108), one should first evaluate

$$\frac{a''}{a} = a^2 \left(H^2 + \frac{\ddot{a}}{a} \right) = a^2 \left(2H^2 + \dot{H} \right) = a^2 H^2 (2 - \varepsilon) = \frac{2 - \varepsilon}{(1 - \varepsilon)^2 \eta^2}. \quad (8.117)$$

We infer that, to first order in ε ,

$$\frac{a''}{a} \simeq \frac{2 + 3\varepsilon}{\eta^2}. \quad (8.118)$$

The Klein–Gordon equation therefore takes the form (8.109) with

$$\nu_\chi^2 = \frac{9}{4} + 3\varepsilon - \frac{m_\chi^2}{H^2}. \quad (8.119)$$

For a light field, the power spectrum then takes the form (8.113), which corresponds to a spectral index

$$n_\chi - 1 \equiv \frac{d \ln \mathcal{P}_\chi}{d \ln k} = 3 - 2\nu_\chi \simeq 2 \left(\frac{m_\chi^2}{3H^2} - \varepsilon \right). \quad (8.120)$$

8.3.3.3 Conclusion

We recover the expected result. The spectrum can be *blue-tilted* ($n_\chi > 1$), *red-tilted* ($n_\chi < 1$) or scale invariant. There is then, respectively, more power in the ultraviolet (small wavelengths) or in the infra-red (long wavelengths). The spectral index gets a contribution from the mass of the test field and from the parameter ε that characterizes the k dependence of H_k .

For a massless scalar field, the spectrum is always red-tilted, independently of the details of the model of inflation. As an example, consider a model of chaotic inflation with inflaton mass m_φ . The spectral index is then

$$n_\chi - 1 = \frac{2}{3H^2} (m_\chi^2 - m_\varphi^2).$$

The nature of the spectrum then depends on the relative masses of both fields.

8.4 Quantum fluctuations of the inflaton

The previous section has cleared the ground for the study of primordial fluctuations generated by the inflaton. Since this field dominates the dynamics of the Universe, we cannot neglect the fluctuations of the geometry and have to follow the gauge invariant relativistic perturbation equations.

8.4.1 Overview of the questions to address and expected results

To start with, any fluctuation of the scalar field will generate metric perturbations. We should first study the coupled inflaton–gravity system to understand how perturbations evolve, mainly after they become super-Hubble. We start from 10 perturbations of the metric and 1 for the scalar field, to which we have to subtract 4 gauge freedoms and 4 constraint equations (2 scalar and 2 vector). We thus expect to identify 3 independent degrees of freedom to describe the full dynamics.

The second question to address is that of the initial conditions and quantization. In the previous examples, we have decided to quantize the field $v = a\chi$, without any justification. Actually, as (8.89) suggests, the Klein–Gordon equation for the field v is that of a harmonic oscillator of variable mass $m^2 = -a''/a$. One can conjecture that it is the canonical variable to quantize, which will be demonstrated. The canonical variables to quantize when gravity is taken into account will have to be identified.

Before we begin, note that if metric perturbations are neglected and if \dot{H} is neglected compared to H^2 , the Klein–Gordon equation for the super-Hubble inflaton perturbations is identical to the time derivative of the background Klein–Gordon equation (8.26)

$$\ddot{\varphi} + 3H\dot{\varphi} + V_{,\varphi\varphi}\dot{\varphi} = 0. \quad (8.121)$$

Qualitatively, we infer that $\delta\varphi = -\dot{\varphi}_0\delta t$, where φ_0 is the homogeneous solution, which implies that $\varphi(\mathbf{x}, t) \sim \varphi_0[t - \delta t(\mathbf{x})]$. Due to the long wavelength perturbations, the number of e-folds between various super-Hubble zones will be modulated. This will in turn give rise to metric fluctuations of the order of $H\delta t(\mathbf{x})$. We infer that $\delta t \sim \delta\varphi/\dot{\varphi} \sim H/2\pi\dot{\varphi}$ so that the metric perturbations are approximately of the order of $H^2/2\pi\dot{\varphi}$. So perturbations are due to the fact the inflaton reaches the bottom of its potential at different times in different points of the Universe. The isodensity or isofield surfaces, are then not constant-time hypersurfaces.

We will show that the inflaton also gives rise to primordial gravitational waves. The typical amplitudes of the density perturbations and gravity waves are of order

$$\frac{\delta\rho}{\rho} \propto \frac{V^{3/2}}{V_{,\varphi}} \sim \frac{H^2}{2\pi\dot{\varphi}}, \quad h \sim \frac{H}{2\pi M_p}.$$

The primordial spectra thus contain information on the dynamics and energy scale of inflation. Let us stress from the outset that the de Sitter limit ($\dot{\varphi} \rightarrow 0$) is singular.

8.4.2 Perturbed quantities

8.4.2.1 Stress-energy tensor

The perturbation of the stress-energy tensor (8.21) of a scalar field is

$$\begin{aligned} \delta T_{\mu\nu} &= 2\partial_{(\nu}\varphi\partial_{\mu)}\delta\varphi - \left(\frac{1}{2}g^{\alpha\beta}\partial_\alpha\varphi\partial_\beta\varphi + V \right) \delta g_{\mu\nu} \\ &\quad - g_{\mu\nu} \left(\frac{1}{2}\delta g^{\alpha\beta}\partial_\alpha\varphi\partial_\beta\varphi + g^{\alpha\beta}\partial_\alpha\delta\varphi\partial_\beta\varphi + V'\delta\varphi \right). \end{aligned} \quad (8.122)$$

Using the SVT decomposition of the metric (5.52), its components take the form

$$\begin{aligned}
\delta T_{00} &= \varphi' \delta \varphi' + 2a^2 V A + a^2 \frac{dV}{d\varphi} \delta \varphi, \\
\delta T_{0i} &= D_i (\varphi' \delta \varphi) + B_i \left(\frac{1}{2} \varphi'^2 - B a^2 V \right), \\
\delta T_{ij} &= \left[\varphi' \delta \varphi' + (C - A) \varphi'^2 - 2a^2 V C - a^2 \frac{dV}{d\varphi} \delta \varphi \right] \gamma_{ij} \\
&\quad + (\varphi'^2 - 2a^2 V) (D_i D_j E + D_{(i} \overline{E}_{j)} + \overline{E}_{ij}). \tag{8.123}
\end{aligned}$$

It is usually more efficient to use the mixed components to obtain the equations,

$$\begin{aligned}
a^2 \delta T_0^0 &= -\varphi' \delta \varphi' - a^2 \frac{dV}{d\varphi} \delta \varphi + A \varphi'^2, \\
a^2 \delta T_i^0 &= -D_i (\varphi' \delta \varphi), \\
a^2 \delta T_0^i &= D^i (\varphi' \delta \varphi + B \varphi'^2) + \overline{B}^i \varphi'^2, \\
a^2 \delta T_j^i &= \left(\varphi' \delta \varphi' - \varphi'^2 A - a^2 \frac{dV}{d\varphi} \delta \varphi \right) \delta_j^i. \tag{8.124}
\end{aligned}$$

8.4.2.2 Gauge invariant variables

Only one quantity appears to describe matter perturbations, namely the perturbation of the scalar field itself. Like any scalar quantity, under a change of coordinates of the form (5.57), it transforms [see (5.80)], to first order, as

$$\delta \varphi \rightarrow \delta \varphi + \varphi' T. \tag{8.125}$$

We can therefore introduce the two following gauge invariant variables

$$\boxed{\chi = \delta \varphi + \varphi' (B - E'),} \tag{8.126}$$

$$\boxed{Q = \delta \varphi - \varphi' \frac{C}{\mathcal{H}},} \tag{8.127}$$

using (5.63). These two quantities are related by

$$Q = \chi + \varphi' \frac{\Psi}{\mathcal{H}}. \tag{8.128}$$

The variable χ , also frequently written as $\delta \varphi^{(\text{gi})}$, is the perturbation of the scalar field in Newtonian gauge, while Q , often called the Mukhanov–Sasaki variable, is the one in flat-slicing gauge.

8.4.3 Perturbation equations

8.4.3.1 Equations in the SVT decomposition

Scalar modes

Following the same approach as for a perfect fluid (Chapter 5), we obtain the following three equations, respectively, for the components (00) , $(0i)$ and the trace of (ij) ,

$$(\Delta + 3K)\Psi - 3\mathcal{H}(\Psi' + \mathcal{H}\Phi) = \frac{\kappa}{2} \left(\varphi'\chi' - \varphi'^2\Phi + a^2 \frac{dV}{d\varphi}\chi \right), \quad (8.129)$$

$$\Psi' + \mathcal{H}\Phi = \frac{\kappa}{2}\varphi'\chi, \quad (8.130)$$

$$\begin{aligned} \Psi'' + 2\mathcal{H}\Psi' - K\Psi + \mathcal{H}\Phi' + (2\mathcal{H}' + \mathcal{H}^2)\Phi + \frac{1}{3}\Delta(\Phi - \Psi) = \\ \frac{\kappa}{2} \left(\varphi'\chi' - \varphi'^2\Phi - a^2 \frac{dV}{d\varphi}\chi \right). \end{aligned} \quad (8.131)$$

The perturbed Klein–Gordon equation then takes the form

$$\chi'' + 2\mathcal{H}\chi' - \Delta\chi + a^2 \frac{d^2V}{d\varphi^2}\chi = 2(\varphi'' + 2\mathcal{H}\varphi')\Phi + \varphi'(\Phi' + 3\Psi'). \quad (8.132)$$

Only two of the four equations (8.129)–(8.132) are independent. The Klein–Gordon equation in terms of the variable Q takes the form

$$Q'' + 2\mathcal{H}Q' - \Delta Q + a^2 \frac{d^2V}{d\varphi^2}Q = \varphi' \left(X' - \frac{1}{\mathcal{H}}\Delta\Psi \right), \quad (8.133)$$

where we recall that $X = \Phi + \Psi + (\Psi/\mathcal{H})'$ [see (5.67)]. Actually, using the Einstein equations, as we shall see later, this equation can be written in a closed form for Q .

Moreover, (8.124) implies that $\delta T_i^j \propto \delta_i^j$, which directly implies that the two Bardeen potentials must be equal,

$$\Phi = \Psi. \quad (8.134)$$

Let us stress one of the properties of these equations in the case of a de Sitter phase where the Universe is dominated by a pure cosmological constant. In this case $\varphi' = 0$ and the Einstein equations imply that

$$\text{de Sitter phase: } \Phi = \Psi = 0. \quad (8.135)$$

This does not mean that the scalar field does not fluctuate but only that its fluctuations do not couple to the geometry and do not generate any metric fluctuations. This explains why the de Sitter limit is often singular in our equations. In what follows, we assume that $\varphi' \neq 0$.

Vector modes

Since a scalar field does not contain any vector sources, there are no vector equations associated with the Klein–Gordon equation and there is only one Einstein equation for these modes

$$\boxed{\bar{\Phi}'_i + 2\mathcal{H}\bar{\Phi}_i = 0.} \quad (8.136)$$

The vector modes are therefore rapidly washed out since $\bar{\Phi}_i \propto a^{-2}$. Actually, the constraint (5.113) reduces, in the case of a scalar field, to

$$(\Delta + 2K)\bar{\Phi}_i = 0.$$

We can conclude, independently of any model of inflation, that *the vector modes are completely absent at the end of inflation*.

Tensor modes

There is only one tensor equation that can be obtained from the Einstein equations

$$\boxed{\bar{E}_{ij}'' + 2\mathcal{H}\bar{E}'_{ij} - (\Delta - 2K)\bar{E}_{ij} = 0.} \quad (8.137)$$

The general study of this equation has been performed in Section 5.3.1 of Chapter 5. As in that chapter, we can expand the gravitational waves into their polarization tensors

$$\bar{E}_{ij}(\mathbf{k}, \eta) = \sum_{\lambda} \bar{E}_{\lambda}(\mathbf{k}, \eta) \varepsilon_{ij}^{\lambda}(\mathbf{k}), \quad (8.138)$$

where $\lambda = +, \times$ indicates the polarization. This symmetric tensor satisfies

$$\varepsilon_{ij}^{\lambda}(\mathbf{k}) \delta^{ij} = 0, \quad k^i \varepsilon_{ij}^{\lambda}(\mathbf{k}) = 0, \quad \varepsilon_{ij}^{\lambda}(\mathbf{k}) \varepsilon_{\lambda'}^{ij}(\mathbf{k}) = \delta_{\lambda\lambda'}, \quad \varepsilon_{ij}^{\lambda}(-\mathbf{k}) = [\varepsilon_{ij}^{\lambda}(\mathbf{k})]^*. \quad (8.139)$$

The last condition is not mandatory but insures that $\bar{E}_{\lambda}(\mathbf{k}, \eta) = \bar{E}_{\lambda}^*(-\mathbf{k}, \eta)$.

8.4.3.2 Other forms of the scalar equations

Despite the apparent complexity of the scalar equations (8.129)–(8.132), using the evolution equation for φ , they can be reduced to a second-order equation with time-dependent coefficients for the gravitational potential Φ ,

$$\Phi'' + 2 \left(\mathcal{H} - \frac{\varphi''}{\varphi'} \right) \Phi' - \left[\Delta - 2 \left(\mathcal{H}' - \mathcal{H} \frac{\varphi''}{\varphi'} - 2K \right) \right] \Phi = 0. \quad (8.140)$$

This equation is valid only if $\varphi' \neq 0$ and therefore cannot be used for a strict de Sitter phase or during the phase where the scalar field oscillates at the bottom of its potential, as during reheating.

Equation (8.140) can take a more suggestive form if we introduce the variables

$$u_s \equiv \frac{2}{\kappa}(\rho + P)^{-1/2}\Phi = \frac{2}{\sqrt{3}\kappa} \frac{a^2\theta}{\mathcal{H}}\Phi, \quad (8.141)$$

with

$$\theta = \frac{1}{a} \left(\frac{\rho}{\rho + P} \right)^{1/2} \left(1 - \frac{3K}{\kappa\rho a^2} \right)^{1/2} = \frac{\mathcal{H}}{a} \left[\frac{2}{3} (\mathcal{H}^2 - \mathcal{H}' + K) \right]^{-1/2}, \quad (8.142)$$

following what was previously done for a perfect fluid in Section 5.3.2 of Chapter 5. It then reduces to the equation

$$u_s'' - \left[\Delta + \frac{\theta''}{\theta} - 3K(1 - c_s^2) \right] u_s = 0, \quad (8.143)$$

with

$$c_s^2 = \frac{P'}{\rho'} \equiv -\frac{1}{3} \left(1 + 2 \frac{\varphi''}{\mathcal{H}\varphi'} \right). \quad (8.144)$$

This equation is identical to (5.147) obtained for a fluid. Note that the definition for u_s is the same as that of Ref. [10] for the case of a fluid but differs by a factor $2/\kappa$ for a scalar field since $u_s = (2/\kappa)a\Phi/\varphi'$. We prefer to keep the same definition for both types of matter.

When $K = 0$, this equation can be interpreted as the equation for a harmonic oscillator with a time-dependent frequency $\omega^2 = k^2 - \theta''/\theta$, or as a Schrödinger equation with potential $U = \theta''/\theta$. This potential contains derivatives of the scale factor up to fourth order.

8.4.4 Evolution of the long-wavelength modes

We restrict our analysis to a Universe with Euclidean spatial sections, $K = 0$, in the rest of this section.

8.4.4.1 Integral solution

For long-wavelength modes, $u_s = \theta$ is an obvious solution of (8.143), whose solutions are given by the integral form

$$u_s(\mathbf{k}, \eta) = \theta(\eta) \left[A(\mathbf{k}) \int \frac{d\eta}{\theta^2(\eta)} + B(\mathbf{k}) \right], \quad (8.145)$$

where $A(\mathbf{k})$ and $B(\mathbf{k})$ are two functions of integration. It can be rewritten as

$$u_s(\mathbf{k}, \eta) = \frac{C(\mathbf{k})}{\varphi'} \left[\frac{1}{a} \int^\eta d\eta a^2(\eta) \right]', \quad (8.146)$$

where the second integration constant has been absorbed in the integral bounds. This equation allows us to conclude that

$$\Phi(\mathbf{k}, \eta) = \frac{\sqrt{3}}{2} C(\mathbf{k}) \left[\frac{1}{a} \int^\eta d\eta a^2(\eta) \right]' = \frac{\sqrt{3}}{2} C(\mathbf{k}) \left(1 - \frac{H}{a} \int adt \right). \quad (8.147)$$

8.4.4.2 Conserved quantity

In Chapter 5, we introduced the quantity ζ defined by (5.149) as

$$\zeta \equiv \Phi + \frac{2}{3} \frac{\Phi' + \mathcal{H}\Phi}{\mathcal{H}(1+w)}.$$

ζ represents the curvature perturbation in comoving gauge. Equation (8.140) is strictly analogous to (5.143) so that ζ satisfies the general equation of conservation (5.150). One can check that, during the period of inflation (assuming $K = 0$), this equation takes the simplified form

$$\frac{3}{2}\mathcal{H}(1+w)\zeta' = \Phi'' + 2\left(\mathcal{H} - \frac{\varphi''}{\varphi'}\right)\Phi' + 2\left(\mathcal{H}' - \mathcal{H}\frac{\varphi''}{\varphi'}\right)\Phi. \quad (8.148)$$

Equation (8.140) thus implies that

$$\zeta' = \frac{2}{3\mathcal{H}(1+w)}\Delta\Phi. \quad (8.149)$$

8.4.4.3 Curvature perturbation

The curvature perturbation of constant density hypersurfaces in flat-slicing gauge (5.153),

$$\zeta_{\text{BST}} = -C + \frac{1}{3}\frac{\delta\rho}{\rho + P},$$

is also a gauge invariant quantity that is also conserved during the evolution. As was shown previously [see (5.156)], this quantity is related to ζ by the relation

$$\zeta = \zeta_{\text{BST}} - \frac{1}{3}\frac{\Delta\Phi}{\mathcal{H}' - \mathcal{H}^2}, \quad (8.150)$$

so that for super-Hubble modes, $\zeta \simeq \zeta_{\text{BST}}$.

In the literature, it is common to find the quantity

$$\mathcal{R} = C - \mathcal{H}\frac{\delta\varphi}{\varphi'} = -\frac{\mathcal{H}}{\varphi'}Q. \quad (8.151)$$

This is the curvature perturbation on a comoving slicing so that \mathcal{R} represents the gravitational potential on constant- φ hypersurfaces. Equation (8.130) can be used to establish that

$$\mathcal{R} = -\zeta \quad \text{if} \quad k\eta \ll 1. \quad (8.152)$$

8.4.4.4 End of inflation

Since ζ is constant, one can relate the value of the gravitational potential on super-Hubble scales during inflation to its value during the radiation era, without worrying about the details of the transition between the two epochs. From (5.158), one concludes that

$$\Phi_{\text{RDU}}(k\eta \ll 1) = \frac{2}{3}\frac{1+\varepsilon}{\varepsilon}\Phi_{\text{inf}}(k\eta \ll 1). \quad (8.153)$$

\mathcal{R} and ζ_{BST} are constant during the transition and the sudden change in the equation of state amplifies the gravitational potential.

In what follows, it will be more convenient to characterize scalar perturbations using one of these conserved quantities as they are not affected by the change in the equation of state, unlike the gravitational potential. ζ is a purely geometric quantity that characterizes density perturbations independently of the exact nature of matter and it allows us to propagate the spectrum of perturbations of the observable modes from inflation to the epoch where we would like to start cosmological computations. In particular, it allows the details of the description of the reheating phase to be avoided.

8.4.5 Junction conditions and their applications

Assuming that the transition between the inflationary phase and the radiation era is instantaneous, one can recover the previous results from junction conditions.

8.4.5.1 Junction conditions

Let us start by determining the junction conditions that shall be satisfied when two space-times are ‘glued’ along a hypersurface Σ . The space-time embedding of this hypersurface into the (four-dimensional) space-time is defined by

$$x^\mu = \bar{x}^\mu(\sigma^a), \quad (8.154)$$

where the σ^a are the intrinsic coordinates and \bar{x}^μ the functions describing the embedding. The induced metric on the hypersurface is then

$$\gamma_{ab} = g_{\mu\nu} \frac{\partial \bar{x}^\mu}{\partial \sigma^a} \frac{\partial \bar{x}^\nu}{\partial \sigma^b}. \quad (8.155)$$

This metric can be seen as a tensor defined in the space-time

$$\bar{\gamma}^{\mu\nu} = \gamma^{ab} \frac{\partial \bar{x}^\mu}{\partial \sigma^a} \frac{\partial \bar{x}^\nu}{\partial \sigma^b}, \quad (8.156)$$

called the *first fundamental form*. The Ricci tensor, R_{ab} , of the induced metric allows us to define

$$\bar{R}^{\mu\nu} = R^{ab} \frac{\partial \bar{x}^\mu}{\partial \sigma^a} \frac{\partial \bar{x}^\nu}{\partial \sigma^b}. \quad (8.157)$$

The tensor $\bar{\gamma}_\nu^\mu$ acts as a projection tensor, projecting tensors at a point of Σ into tangential tensors and can be used to define the orthogonal projection operator by

$$\perp_\nu^\mu = \delta_\nu^\mu - \bar{\gamma}_\nu^\mu. \quad (8.158)$$

These definitions actually do not depend on the codimension of Σ . In the case of a hypersurface (i.e. codimension 1), there is a unique perpendicular direction so that $\perp_\nu^\mu = n^\mu n_\nu$, where n^μ is the unit vector normal to Σ (satisfying $n_\mu n^\mu = 1$ and $\bar{\gamma}_{\mu\nu} n^\nu = 0$) and we have

$$\bar{\gamma}_{\mu\nu} = g_{\mu\nu} - n_\mu n_\nu. \quad (8.159)$$

We can then define the *extrinsic curvature tensor*

$$K_{\mu\nu} = -\bar{\gamma}_\nu^\sigma \bar{\gamma}_\mu^\alpha \nabla_\alpha n_\sigma. \quad (8.160)$$

If the vector field n^μ , initially defined only on Σ , is geodesically extended in the neighbourhood of the hypersurface (with $n^\alpha \nabla_\alpha n_\mu = 0$) then

$$K_{\mu\nu} = -\nabla_\mu n_\nu. \quad (8.161)$$

One can show [17] in a general way, that two space-times can be glued along Σ if

$$[\bar{\gamma}_{\mu\nu}]_\pm = 0, \quad [K_{\mu\nu} - K\bar{\gamma}_{\mu\nu}]_\pm = \kappa\tau_{\mu\nu}, \quad (8.162)$$

where A_+ and A_- represent the values of A evaluated on each side of the hypersurface and $[A]_\pm \equiv A_+ - A_-$ is the jump of A across the hypersurface.

The first equation implies that the induced metric is the same whether it is defined from one side or the other of Σ . In the second expression, $\tau_{\mu\nu}$ is the stress-energy tensor of matter fields localized on the hypersurface, including its tension when seen as a membrane. Extensions of the description of such codimension 1 membranes are detailed in Ref. [18] and the dynamics of a hypersurface will be described in Chapter 13 in the case of ‘brane’ cosmology.

8.4.5.2 Application in the context of cosmology

For a cosmological transition, Σ is usually defined by

$$\Sigma = \{q = \text{const.}\}, \quad (8.163)$$

where q is a scalar function characterizing the hypersurface (i.e. $q = \rho$ for a constant-density hypersurface, or $q = \varphi$ for a constant-field hypersurface, etc.) The unit vector normal to Σ is then given by

$$n_\mu = \frac{\partial_\mu q}{\sqrt{-\partial_\alpha q \partial^\alpha q}}, \quad (8.164)$$

and we assume that the two space-times have metrics

$$ds_\pm^2 = a_\pm^2(\eta_\pm) \left\{ -(1 + 2\Phi_\pm) d\eta_\pm^2 + [(1 - 2\Psi_\pm)\gamma_{ij} + h_{ij}^\pm] dx^i dx^j \right\}.$$

The induced metric and the extrinsic curvature tensor are explicitly computed in Appendix C. We find

$$\gamma_b^a = a^2 \left\{ \left[1 - 2 \left(\Psi + \frac{\delta q}{q'} \right) \right] \delta_b^a + h_b^a \right\}, \quad (8.165)$$

and

$$K_b^a = \frac{1}{a} \left\{ -\mathcal{H} \delta_b^a + \left[\mathcal{H}\Phi + \Psi' + (\mathcal{H}' - \mathcal{H}^2) \frac{\delta q}{q'} \right] \delta_b^a + \frac{1}{2} h_b'^a + \partial^a \partial_b \frac{\delta q}{q'} \right\}, \quad (8.166)$$

where δq is the perturbation of q in Newtonian gauge. The junction conditions (8.162) then take the form [19, 20]

$$\text{background metric : } [a]_{\pm} = 0, \quad [\mathcal{H}]_{\pm} = 0, \quad (8.167)$$

$$\text{scalar modes : } [\Psi]_{\pm} = 0, \quad \left[\frac{\delta q}{q'} \right]_{\pm} = 0, \quad \left[\Psi' + \mathcal{H}\Phi + \mathcal{H}'\frac{\delta q}{q'} \right]_{\pm} = 0, \quad (8.168)$$

$$\text{tensor modes : } [h_{ij}]_{\pm} = 0, \quad [h'_{ij}]_{\pm} = 0. \quad (8.169)$$

The first condition for the background metric tells us that the scale factor and expansion rate must be continuous. But note that the equation of state can be discontinuous, and so \mathcal{H}' as well.

To apply these conditions for perturbations, we should assume that the transition is fast compared to the typical oscillation time of the modes considered, which amounts to assuming that these mode are super-Hubble at the time of the transition. Note also that even if this result was obtained by assuming that $K = 0$, it is possible to generalize it for any value of K [21].

8.4.5.3 End of inflation

Let us apply these junction conditions to relate the end of inflation to the radiation era. Relation (8.167) allows us to conclude that the scale factors then take the form

$$a_-(\eta_-) = \frac{C}{(-\eta_-)^{1+\varepsilon}}, \quad a_+(\eta_+) = \frac{C}{[(1+\varepsilon)\eta_*^{1+\varepsilon}]} \left(\frac{\eta_+}{\eta_*} \right),$$

if the transition happens at $\eta_+ = \eta_* = -(1-\varepsilon)\eta_-$.

Assuming that the transition occurs on a constant-density hypersurface, $q = \rho$ and the conditions (8.168) imply that $[\zeta_{\text{BST}}]_{\pm} = 0$ and so $[\zeta + \Delta\Phi/3(\mathcal{H}' - \mathcal{H}^2)]_{\pm} = 0$. Using (5.117), we infer that for super-Hubble modes,

$$\left[\frac{3\mathcal{H}(\Phi' + \mathcal{H}\Phi)}{1+w} \right]_{\pm} = 0. \quad (8.170)$$

Note that if the transition surface had been chosen as a constant-field hypersurface, then we would have concluded $[\mathcal{R}]_{\pm} = 0$ directly. For super-Hubble modes, it is therefore equivalent to assuming that the end of inflation is a constant-field or a constant-density hypersurface.

In both eras, the gravitational potential is given by $\Phi_- = A_- + B_-(-\eta_-)$ and $\Phi_+ = A_+ + B_+/\eta_+^3$, respectively. The junction conditions imply that

$$A_+ = \frac{2}{3} \frac{1+\varepsilon}{\varepsilon} A_-, \quad B_+ = \frac{1}{3} \frac{\varepsilon-2}{\varepsilon} A_+, \quad (8.171)$$

once the decaying mode during inflation has been neglected. Thus, we recover the previous result: $\Phi_+ \sim [2(1+\varepsilon)/2\varepsilon]\Phi_-$. However, note that the growing mode inherits a contribution from the gravitational potential during inflation (Φ_-) and a contribution from the density perturbations at the time of the transition $[(2-\varepsilon)\Phi_-/\varepsilon]$.

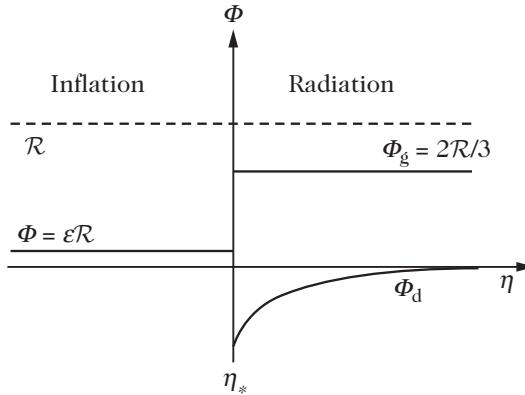


Fig. 8.10 Junction between the inflationary phase and the post-inflationary radiation era for the super-Hubble modes. ζ and \mathcal{R} are constant during the transition so that the amplification of the gravitational mode is associated with the appearance of a decaying mode, Φ_d , that can then be neglected at late times.

Figure 8.10 illustrates how the growing mode of the gravitational potential can be amplified during inflation while keeping the curvature perturbation constant, at the price of exciting the decaying mode. Let us stress that in the de Sitter limit, $\varepsilon \rightarrow 0$ and thus $\Phi \rightarrow 0$. In this case, perturbations of the post-inflationary era come entirely from perturbations of the surface of transition.

8.4.6 Quantization of the density perturbations

The previous section detailed the evolution of perturbations but did not discuss the choice of the initial state of the fields. To follow the example of the test scalar field, we should determine which are the variables to quantize. This aspect is detailed in Ref. [10] and we only recall the main steps of the construction.

8.4.6.1 Canonical variable

To determine the canonical variables, let us first introduce the variable

$$v = aQ = -z\mathcal{R}, \quad (8.172)$$

where Q and \mathcal{R} are the scalar field perturbation in the flat-slicing gauge and the curvature perturbation in comoving gauge, and where z is defined by

$$z \equiv \frac{a\dot{\varphi}}{H} = \frac{a\varphi'}{\mathcal{H}}. \quad (8.173)$$

The evolution equations for the perturbations have been obtained by a variational principle from the Einstein–Hilbert action and a minimally coupled scalar field. Since these equations are linear, they can be obtained from the same actions expanded to

quadratic order. Starting from the expressions (8.172) and (8.173), one can show that the total action takes the form

$$\delta^{(2)}S = \frac{1}{2} \int d\eta d^3x \left[(v')^2 - \delta^{ij} \partial_i v \partial_j v + \frac{z''}{z} v^2 \right] \equiv \int \mathcal{L} d^4x, \quad (8.174)$$

up to terms involving a total derivative and that do not contribute to the equations of motion.

This rather technical calculation is performed explicitly in Refs. [10, 22] (see also Ref. [84] for a generalization to Bianchi space-times). Deriving the evolution equations to first order is strictly equivalent to varying the action expanded to second order, however, the second approach gives us more information on the structure of the theory. This action is that of a scalar field with a time-dependent mass, $m^2 = -z''/z$ in a Minkowski space-time. It is interesting to stress that the natural variable is neither Φ nor ζ but v . Note that if we had started from the evolution equations [see, e.g., (8.143)], we could have wrongly concluded that u_s was the canonical variable. This equivalence implies that we can quantize this theory as we would quantize a scalar field evolving in a time-dependent exterior field [23], where here the time dependence would find its origin in the space-time dynamics [24]. Note also the similarity with the case of a test scalar field in an expanding space-time, as considered in Section 8.3.

It is therefore the field v that should be promoted to the status of quantum operator in second quantization, (see Chapter 2)

$$\hat{v}(\mathbf{x}, \eta) = \int \frac{d^3k}{(2\pi)^{3/2}} \left[v_k(\eta) e^{i\mathbf{k}\cdot\mathbf{x}} \hat{a}_k + v_k^*(\eta) e^{-i\mathbf{k}\cdot\mathbf{x}} \hat{a}_k^\dagger \right]. \quad (8.175)$$

In what follows we will work in the Heisenberg representation, in which the time-dependence is carried by the operators.

The first step in performing a canonical quantization is to introduce the conjugate momentum of v ,

$$\pi = \frac{\delta \mathcal{L}}{\delta v'} = v', \quad (8.176)$$

which is also promoted to the status of operator, $\hat{\pi}$. We can then get the Hamiltonian

$$H = \int (v' \pi - \mathcal{L}) d^4x = \frac{1}{2} \int \left(\pi^2 + \delta^{ij} \partial_i v \partial_j v - \frac{z''}{z} v^2 \right) d^4x. \quad (8.177)$$

The operators \hat{v} and $\hat{\pi}$ have to satisfy canonical commutation relations on constant time hypersurfaces

$$[\hat{v}(\mathbf{x}, \eta), \hat{v}(\mathbf{y}, \eta)] = [\hat{\pi}(\mathbf{x}, \eta), \hat{\pi}(\mathbf{y}, \eta)] = 0, \quad [\hat{v}(\mathbf{x}, \eta), \hat{\pi}(\mathbf{y}, \eta)] = i\delta^{(3)}(\mathbf{x} - \mathbf{y}). \quad (8.178)$$

The equation of motion for v obtained from the Lagrangian (8.174) provides the field equation for the operator \hat{v} , which is similar to the Klein–Gordon equation in flat space-time,

$$\hat{v}'' - \left(\Delta + \frac{z''}{z} \right) \hat{v} = 0.$$

(8.179)

Indeed, this is also the classical Klein–Gordon equation that can be derived from (8.133) after performing the change of variables (8.172) from Q to v . This equation is actually equivalent to the Heisenberg equations

$$\hat{v}' = i [\hat{H}, \hat{v}], \quad \hat{\pi}' = i [\hat{H}, \hat{\pi}], \quad (8.180)$$

where \hat{H} is simply the Hamiltonian (8.177) in terms of \hat{v} and $\hat{\pi}$.

As for quantization in Minkowski space-time (see Chapter 2), the creation and annihilation operators appearing in the decomposition (8.175) satisfy the standard commutation rules

$$[\hat{a}_{\mathbf{k}}, \hat{a}_{\mathbf{p}}] = [\hat{a}_{\mathbf{k}}^\dagger, \hat{a}_{\mathbf{p}}^\dagger] = 0, \quad [\hat{a}_{\mathbf{k}}, \hat{a}_{\mathbf{p}}^\dagger] = \delta^{(3)}(\mathbf{k} - \mathbf{p}). \quad (8.181)$$

These commutation rules are consistent with the commutation rules (8.178) only if the field v_k is normalized according to

$$W(k) \equiv v_k v'_k{}^* - v_k^* v'_k = i, \quad (8.182)$$

since

$$[\hat{v}(\mathbf{x}, \eta), \hat{\pi}(\mathbf{y}, \eta)] = \int \frac{d^3 k}{(2\pi)^3} e^{i\mathbf{k}\cdot(\mathbf{x}' - \mathbf{x})} W(k).$$

This determines the normalization of their Wronskian, W .

Then, one needs to construct the Fock representation of the Hilbert space on which the operators \hat{v} and $\hat{\pi}$ act. In Minkowski space-time, the vacuum state is uniquely defined by the condition $\hat{a}_{\mathbf{k}} |0\rangle = 0$ for all \mathbf{k} and the other states are obtained by repeated action of the creation operator $\hat{a}_{\mathbf{k}}^\dagger$. This amounts to keeping only negative-frequency modes.

The vacuum state $|0\rangle$ is thus defined by the condition that this state is annihilated by all the operators $\hat{a}_{\mathbf{k}}$,

$$\forall \mathbf{k}, \quad \hat{a}_{\mathbf{k}} |0\rangle = 0. \quad (8.183)$$

The sub-Hubble modes, i.e. the high-frequency modes compared to the expansion of the Universe, must behave as in a flat space-time. So the asymptotic condition to impose for a state in the vacuum is

$$v_k(\eta) \rightarrow \frac{1}{\sqrt{2k}} e^{-ik\eta}, \quad k\eta \rightarrow -\infty. \quad (8.184)$$

Let us note that in an expanding space-time, the notion of time is fixed by the background evolution that provides a preferred direction. The notion of positive and negative frequency is not time invariant, which implies that during the evolution time of positive frequency will be generated.

Note that this quantization procedure amounts to treating gravity in a quantum way to linear order since the gravitational potential is promoted to the status of operator.

8.4.6.2 Particle creation

If we pick up an arbitrary initial time η_i then

$$v_k(\eta_i) = \frac{1}{\sqrt{\omega(\eta_i, k)}}, \quad v'_k(\eta_i) = i\sqrt{\omega(\eta_i, k)}, \quad (8.185)$$

defines the positive frequency solutions at η_i , indeed for oscillating modes. We can thus define a vacuum state, $|0_i\rangle$ following the previous procedure.

At a later time $\eta_1 > \eta_i$, the modes defined by (8.185) will no longer be positive frequency and an observer at η_1 will define another vacuum state $|0_1\rangle$ satisfying $\hat{c}_k|0_1\rangle = 0$ for all \mathbf{k} , \hat{c}_k being the annihilation operators for the positive-frequency modes at η_1 .

Since the mode equation is linear, it is clear that the positive-frequency modes at η_1 , w_k are related to mode functions v_k by a linear transformation

$$w_k = \alpha_{\mathbf{k}} v_k + \beta_{\mathbf{k}} v_k^*,$$

with the two complex coefficients satisfying $|\alpha_{\mathbf{k}}|^2 - |\beta_{\mathbf{k}}|^2 = 1$. The number of particles of mode \mathbf{k} at η_1 will thus be defined by $N_{\mathbf{k}}^{(1)} = \hat{c}_{\mathbf{k}}^\dagger \hat{c}_{\mathbf{k}}$. Since the two sets of creation and annihilation operators are related by

$$\hat{a}_{\mathbf{k}} = \alpha_{\mathbf{k}} \hat{c}_{\mathbf{k}} + \beta_{\mathbf{k}}^* \hat{c}_{\mathbf{k}}^\dagger,$$

implying that $\hat{c}_{\mathbf{k}} = \alpha_{\mathbf{k}}^* \hat{c}_{\mathbf{k}} - \beta_{\mathbf{k}}^* \hat{a}_{\mathbf{k}}^\dagger$, we deduce that the state $|0_i\rangle$ is seen as a state containing $\langle 0_i | N_{\mathbf{k}}^{(1)} | 0_i \rangle = |\beta_{\mathbf{k}}|^2$ particles by the observer at η_1 .

This illustrates the time dependence of the notion of positive frequency and vacuum state. Fortunately the ambiguity of the definition of the vacuum does not make important differences for small wavelength modes for which the WKB approximation holds. This mechanism of quantum particle creation is discussed in details in Section 10.2 of Chapter 10.

8.4.6.3 Relation with the perturbation variables

The previous definitions allow us to conclude that

$$z = \sqrt{\frac{3}{\kappa}} \theta^{-1} \quad \text{and} \quad \Delta u_s = -z(v/z)' . \quad (8.186)$$

Einstein's equations imply that the gravitational potential is given by

$$\Delta \Phi = \frac{\kappa}{2} \frac{\varphi'^2}{\mathcal{H}} \left(\frac{v}{z} \right)' \quad \text{and} \quad \left(\frac{a^2 \Phi}{\mathcal{H}} \right)' = \frac{\kappa}{2} z v, \quad (8.187)$$

and we also have the useful relation

$$2\mathcal{H}X = \frac{\kappa}{a} \varphi' v, \quad (8.188)$$

where X has been defined in (5.67).

Following the same procedure as for a test scalar field considered in Section 8.3, (8.172) implies that the power spectrum of the metric perturbations is

$$\boxed{\mathcal{P}_{\mathcal{R}} = \mathcal{P}_\zeta = \frac{k^3}{2\pi^2} \left| \frac{v_k}{z} \right|^2,} \quad (8.189)$$

using the definitions of Section B.5 of Appendix B for the normalization of the power spectrum.

The solution of (8.179) with initial conditions (8.184) allows us to determine the gravitational potential and the power spectrum of \mathcal{R} and ζ . The analysis of Section 8.4.5 then allows us to determine the curvature perturbation and the gravitational potential at the beginning of the post-inflationary phase.

8.4.7 Gravitational waves

8.4.7.1 Quantization

The same procedure can be followed for gravitational waves. Restricting ourselves to tensor degrees of freedom, the action at second order in perturbations takes the form

$$\delta^{(2)}S = \frac{M_p^2}{64\pi} \int a^2(\eta) \left[(h'_{ij})^2 - (D_k h_{ij})^2 - 2K(h_{ij})^2 \right] \sqrt{\gamma} d^3x d\eta. \quad (8.190)$$

Unsurprisingly, the variation of this action gives the equation of propagation (8.137). Using the decomposition (8.138), and defining

$$\mu_\lambda(\mathbf{k}, \eta) = \sqrt{\frac{M_p^2}{8\pi}} a(\eta) \overline{E}_\lambda(\mathbf{k}, \eta) = \sqrt{\frac{M_p^2}{32\pi}} a(\eta) h_\lambda(\mathbf{k}, \eta) \quad (8.191)$$

the action at second order takes the form

$$\delta^{(2)}S = \frac{1}{2} \sum_\lambda \int \left[(\mu'_\lambda)^2 - \gamma^{ij} \partial_i \mu_\lambda \partial_j \mu_\lambda - 2K\mu_\lambda^2 + \frac{a''}{a} \mu_\lambda^2 \right] \sqrt{\gamma} d^3x d\eta. \quad (8.192)$$

In this form, the action for the gravitational waves reduces to the action for two independent scalar fields, μ_λ , with time-dependent mass $m^2 = -a''/a$ evolving in a static space-time. We can therefore proceed as for the scalar perturbations and quantize these two degrees of freedom. In the Heisenberg representation, we promote μ_λ to the status of quantum operator

$$\hat{\mu}_\lambda(\mathbf{x}, \eta) = \int \frac{d^3k}{(2\pi)^{3/2}} \left[\mu_{k,\lambda}(\eta) e^{i\mathbf{k}\cdot\mathbf{x}} \hat{a}_{\mathbf{k},\lambda} + \mu_{k,\lambda}^*(\eta) e^{-i\mathbf{k}\cdot\mathbf{x}} \hat{a}_{\mathbf{k},\lambda}^\dagger \right]. \quad (8.193)$$

If the modes $\mu_{\mathbf{k},\lambda}$ satisfy the normalization (8.182), then the creation and annihilation operators satisfy the commutation relations

$$[\hat{a}_{\mathbf{k},\lambda}, \hat{a}_{\mathbf{p},\lambda'}] = [\hat{a}_{\mathbf{k},\lambda}^\dagger, \hat{a}_{\mathbf{p},\lambda'}^\dagger] = 0, \quad [\hat{a}_{\mathbf{k},\lambda}, \hat{a}_{\mathbf{p},\lambda'}^\dagger] = \delta_{\lambda\lambda'} \delta^{(3)}(\mathbf{k} - \mathbf{p}), \quad (8.194)$$

and for each polarization, the vacuum is defined by

$$\forall \mathbf{k}, \quad \hat{a}_{\mathbf{k},\lambda} |0\rangle = 0, \quad (8.195)$$

and we set the initial conditions in the same way as for scalar perturbations by imposing

$$\mu_{k,\lambda}(\eta) \rightarrow \frac{1}{\sqrt{2k}} e^{-ik\eta}, \quad k\eta \rightarrow -\infty. \quad (8.196)$$

The evolution equation for $\mu_{k,\lambda}$, that we will write simply as μ_k when there is no ambiguity, is

$$\boxed{\mu_k'' + \left(k^2 + 2K - \frac{a''}{a} \right) \mu_k = 0.} \quad (8.197)$$

As previously, this equation can be solved by using the WKB approximation (8.102) with

$$\omega_\lambda^2(k, \eta) = k^2 + 2K - \frac{a''}{a}.$$

8.4.7.2 Power spectrum

The power spectrum of gravitational waves is obtained by solving the evolution equation (8.197) with initial conditions (8.184). This allows us to define the amplitude of super-Hubble tensor modes, which remains constant.

The power spectrum, \mathcal{P}_T , of the tensor modes is defined from the variable h_{ij} and is related to the power spectrum \mathcal{P}_E of one polarization \bar{E}_λ by

$$\mathcal{P}_T = 8\mathcal{P}_E,$$

where the factor 8 arises from the sum over polarizations. Because of isotropy, the two polarizations have the same spectrum \mathcal{P}_E defined by

$$\langle \bar{E}_\lambda(\mathbf{k}, \eta_f) \bar{E}_{\lambda'}^*(\mathbf{k}', \eta_f) \rangle = \frac{2\pi^2}{k^3} \mathcal{P}_T(k) \delta^{(3)}(\mathbf{k} - \mathbf{k}') \delta_{\lambda\lambda'}, \quad (8.198)$$

with

$$\boxed{\mathcal{P}_E = \frac{k^3}{2\pi^2} \frac{8\pi}{M_P^2} \left| \frac{\mu_k}{a} \right|^2}, \quad (8.199)$$

for each polarization. Thus, we conclude that

$$\boxed{\mathcal{P}_T = \frac{k^3}{2\pi^2} \frac{64\pi}{M_P^2} \left| \frac{\mu_k}{a} \right|^2}, \quad (8.200)$$

where each polarization λ contributes in an identical way.

8.4.7.3 Gravitational wave density

The variation of the action (8.190) with respect to the metric allows us to define the stress-energy tensor for the gravitational waves

$$t_{\mu\nu} = \frac{1}{16\pi G_N} \sum_{\lambda} (\partial_{\mu} \bar{E}_{\lambda} \partial_{\nu} \bar{E}_{\lambda} - \bar{g}_{\mu\nu} \partial_{\alpha} \bar{E}_{\lambda} \partial^{\alpha} \bar{E}_{\lambda}). \quad (8.201)$$

The gravitational wave amplitude at any time after the inflationary phase can be decomposed as

$$\bar{E}_{\lambda}(k, \eta) = T(k, \eta) \bar{E}_{\lambda}(k, \eta_f) \quad (8.202)$$

where $T(k, \eta)$ is a transfer function obtained by solving (5.109). The gravitational wave density can then be defined as $\rho_{\text{GW}} = -\langle t_0^0(\mathbf{x}, \eta) \rangle$. As soon as the mode becomes sub-Hubble, it oscillates and one can average ρ_{GW} over a few periods to extract its amplitude. This procedure is well defined as long as the amplitude varies slowly with respect to the wave period. In practice, this amounts to replacing $T(k, \eta)$ by $\bar{T}(k, \eta)$, which is the average value of T over a few periods. We then obtain the energy density per decade

$$\frac{d\rho_{\text{GW}}}{d \ln k} = \frac{1}{8\pi G_N} \left(\frac{k}{a} \right)^2 \mathcal{P}_T(k) \bar{T}(k, \eta), \quad (8.203)$$

or equivalently in terms of the density parameter,

$$\frac{d\Omega_{\text{GW}}}{d \ln k} = \frac{1}{3} \left(\frac{k}{aH} \right)^2 \mathcal{P}_T(k) \bar{T}(k, \eta). \quad (8.204)$$

As long as the mode is super-Hubble, its amplitude is constant (see Chapter 5) and then it enters a damped oscillation regime on sub-Hubble scales. The damping between the time where the mode becomes super-Hubble and today is typically of order $\bar{T}(k, \eta) \simeq (a_k/a_0)^2$ where a_k is the value of the scale factor when $k = aH$. We infer that the gravitational wave spectrum today is

$$\frac{d\Omega_{\text{GW}}}{d \ln k} \propto k^2 a_k^2 \mathcal{P}_T(k). \quad (8.205)$$

If the mode becomes sub-Hubble during the matter-dominated era, $a_k \propto k^{-2}$, whereas during the radiation era $a_k \propto k^{-1}$. We infer that

$$\frac{d\Omega_{\text{GW}}}{d \ln k} \propto \mathcal{P}_T(k) \times \begin{cases} k^{-2} & k < k_{\text{eq}} \\ k^0 & k > k_{\text{eq}} \end{cases}. \quad (8.206)$$

This contribution is represented in Fig. 11.15 for a primordial scale invariant spectrum, $\mathcal{P}_T \propto k^0$.

8.5 Perturbations in the slow-roll regime

We will now apply the previous formalism to inflation in the slow-roll regime. In this approximation, one only needs to determine the functions $z(\eta)$ and $a(\eta)$ that then allow for the integration of (8.179) and (8.197). It is useful to note already at this early stage that

$$z = a \frac{M_p}{\sqrt{4\pi}} \sqrt{\varepsilon}. \quad (8.207)$$

8.5.1 An exact solution: power-law inflation

Before studying the general case, it is useful to consider the case of power-law inflation (8.78) for which the slow-roll parameters remain constant, $\varepsilon = \delta = 1/p$ and $\xi = 0$. In this case, the expression (8.116) is exact, so that

$$aH = \mathcal{H} = -\frac{1}{1-\varepsilon}\frac{1}{\eta}, \quad (8.208)$$

and, as for a de Sitter space-time, η varies between $-\infty$ and 0.

8.5.1.1 Scalar modes

The conformal time evolution of the scale factor (8.76) and scalar field (8.77) are determined explicitly. The variable z , defined by (8.173), is then given by

$$z = \sqrt{p\pi}\eta^{1+\beta},$$

so that

$$\frac{z''}{z} = \frac{\beta(\beta+1)}{\eta^2}. \quad (8.209)$$

Equation (8.179) then takes the form of a Bessel equation

$$v_k'' + \left(k^2 - \frac{\nu^2 - 1/4}{\eta^2} \right) v_k = 0, \quad \nu = -\beta - \frac{1}{2} = \frac{3}{2} + \frac{1}{p-1}. \quad (8.210)$$

The solution of this equation that satisfies the initial condition (8.184) can be expressed in terms of a Hankel function as

$$v_k(\eta) = \frac{\sqrt{\pi}}{2} e^{i(\nu+1/2)\pi/2} \sqrt{-\eta} H_\nu^{(1)}(-k\eta). \quad (8.211)$$

For super-Hubble modes ($k/\mathcal{H} \ll 1$), this solution tends towards

$$v_k(\eta) \rightarrow 2^{\nu-3/2} \frac{\Gamma(\nu)}{\Gamma(3/2)} e^{i(\nu-1/2)\pi/2} \frac{1}{\sqrt{2k}} (-k\eta)^{-\nu+1/2}. \quad (8.212)$$

The expression (8.189) allows us to find that the power spectrum of curvature perturbations is given by

$$\mathcal{P}_\zeta = \frac{H^2}{\pi M_p^2 \varepsilon} \left[2^{\nu-3/2} \frac{\Gamma(\nu)}{\Gamma(3/2)} \right]^2 \left(\nu - \frac{1}{2} \right)^{-2\nu+1} \left(\frac{k}{aH} \right)^{-2\nu+3}. \quad (8.213)$$

The coefficient $(\nu - 1/2)^{-2\nu+1}$ comes from the contribution $(1-\varepsilon)^{2\nu-1}$ that appears when replacing $k\eta$ using (8.208). If we evaluate this spectrum at $aH = k$, it can then be rewritten as

$$\mathcal{P}_\zeta = \frac{1}{\pi M_p^2} \left[2^{\nu-3/2} \frac{\Gamma(\nu)}{\Gamma(3/2)} \right] \left(\nu - \frac{1}{2} \right)^{-2\nu+1} \left(\frac{H^2}{\varepsilon} \right)_{k=aH}, \quad (8.214)$$

which hides the k dependence in the term H^2/ε evaluated at the time where the mode k becomes super-Hubble.

8.5.1.2 Gravitational waves

The study of gravitational waves is in this case identical to that of the scalar modes since in this particular model of inflation $z \propto a$. Equation (8.197) is thus identical to (8.179) and the solution for μ_k is given by (8.211). We infer from (8.200) that

$$\mathcal{P}_T = \frac{64\pi}{M_P^2} \left(\frac{z}{a}\right)^2 \mathcal{P}_\zeta.$$

The gravitational wave power spectrum is thus given by

$$\mathcal{P}_T = 16\varepsilon \mathcal{P}_\zeta.$$

(8.215)

8.5.1.3 Spectral indices

Note that the amplitude of gravitational waves is determined only by the energy scale, H , during the slow-roll regime, while that of the scalar modes depends on both H and the slow-roll parameter ε . The spectral indices, defined by

$$n_s - 1 \equiv \frac{d \ln \mathcal{P}_\zeta}{d \ln k}, \quad n_T \equiv \frac{d \ln \mathcal{P}_T}{d \ln k}, \quad (8.216)$$

are, respectively, given by

$$n_s = -2\nu + 4 = 2\beta + 5 \quad n_T = -2\nu + 3 = 2\beta + 4. \quad (8.217)$$

Note that when $\varepsilon \ll 1$, i.e. when $p \gg 1$, then $n_T = -2/(p-1) \sim -2/p = -2\varepsilon$ to first order in the slow-roll parameters. We deduce the consistency relation to lowest order in the slow-roll parameters

$$\frac{\mathcal{P}_T}{\mathcal{P}_\zeta} = -8n_T. \quad (8.218)$$

8.5.2 General case

In the general case, the slow-roll parameters are time dependent and the evolution equations cannot be integrated exactly. We should thus construct an expansion scheme allowing us to approximate the solution.

8.5.2.1 Expansion scheme

The case of power-law inflation that we considered previously corresponds to constant slow-roll parameters. In an arbitrary model, these parameters vary and their dynamics is dictated, to lowest order, by (8.46) and (8.179), and (8.197) cannot be solved analytically.

In order to make an expansion in terms of the slow-roll parameters, we use (8.46) to replace the time dependence of ε to leading order in all equations, in order to identify

the negligible terms. Relation (8.116) can then be generalized by performing successive integration by parts and expanding ε as a function of η , namely

$$\eta = \int \frac{da}{a\mathcal{H}} = -\frac{1}{\mathcal{H}} + \int \frac{\varepsilon da}{a\mathcal{H}} = -\frac{1}{\mathcal{H}} + \varepsilon \int \frac{da}{a\mathcal{H}} - \int \frac{\varepsilon' da}{a\mathcal{H}^2}, \quad (8.219)$$

to obtain

$$\eta = -\frac{1}{\mathcal{H}} \left[\frac{1}{1-\varepsilon} - 2\varepsilon(\varepsilon - \delta) \right] + \dots \quad (8.220)$$

The parameter ε can now have an arbitrary time dependence. Assuming that ε and δ are small parameters, then to first order we can consider them to be constant (since ε' is of second order) so that

$$\eta = -\frac{1}{\mathcal{H}}(1+\varepsilon) + \mathcal{O}(2). \quad (8.221)$$

We will distinguish the *leading-order* expansion in which only the lowest powers of the slow-roll parameters are conserved from the *next-order* expansion that contains the next to leading-order corrections.

The computation of the power spectrum at next order was initially performed by Stewart and Lyth [25] and the extension to second order can be found in Ref. [26].

8.5.2.2 Expansion of the functions z''/z and a''/a

The two functions z''/z and a''/a are essential for solving (8.179) and (8.197).

Starting from (8.173), we obtain

$$\frac{z'}{z} = \mathcal{H} + \frac{\varphi''}{\varphi'} - \frac{\mathcal{H}'}{\mathcal{H}} = \mathcal{H}(1+\varepsilon-\delta),$$

where the second equality uses relations (8.48). After differentiating this equation and using relations (8.46) and (8.48), we find that

$$\begin{aligned} \frac{z''}{z} &= \mathcal{H}^2 [2 + 2\varepsilon - 3\delta + \delta^2 - \varepsilon\delta + \xi], \\ &= \mathcal{H}^2 [2 + 2\varepsilon - 3\delta + 2\varepsilon^2 + \delta^2 - 4\varepsilon\delta + \xi_H^2], \end{aligned} \quad (8.222)$$

this result being exact. In an identical way, since $a''/a = \mathcal{H}' + \mathcal{H}^2$, we obtain

$$\frac{a''}{a} = (2-\varepsilon)\mathcal{H}^2. \quad (8.223)$$

Relation (8.221) then allows us to express these quantities in terms of the conformal time. To first order, we obtain

$$\frac{z''}{z} = \frac{2 + 6\varepsilon - 3\delta}{\eta^2} + \mathcal{O}(2), \quad \frac{a''}{a} = \frac{2 + 3\varepsilon}{\eta^2} + \mathcal{O}(2).$$

(8.224)

8.5.2.3 Scalar modes

As in the case of power-law inflation, (8.179) reduces to a Bessel equation of the form (8.210) with

$$\nu = \frac{3}{2} + 2\varepsilon - \delta. \quad (8.225)$$

The solution that satisfies the initial conditions (8.184) is given by (8.211). For super-Hubble modes, we should then expand this solution, keeping the leading-order term and the next-order ones. Using $\Gamma(a+h)/\Gamma(a) \sim 1 + h\psi(a)$ and $2^h \sim 1 + h \ln 2$ for $h \ll 1$, we obtain

$$\mathcal{P}_\zeta = \frac{1}{\pi} \frac{H^2}{M_p^2 \varepsilon} [1 - 2(2C+1)\varepsilon + 2C\delta] \left(\frac{k}{aH} \right)^{2\delta-4\varepsilon}, \quad (8.226)$$

where $C = \gamma_E + \ln 2 - 2$, and $\gamma_E \sim 0.5772$ the Euler constant. The spectrum in the slow-roll regime takes the two equivalent forms,

$$\boxed{\mathcal{P}_\zeta = \frac{1}{\pi} \frac{H^2}{M_p^2 \varepsilon} \left[1 - 2(2C+1)\varepsilon + 2C\delta + (2\delta - 4\varepsilon) \ln \left(\frac{k}{aH} \right) \right]}, \quad (8.227)$$

$$\boxed{= \frac{1}{\pi} [1 - 2(2C+1)\varepsilon + 2C\delta] \left(\frac{H^2}{M_p^2 \varepsilon} \right)_{k=aH}.} \quad (8.228)$$

8.5.2.4 Gravitational waves

As in the case of power-law inflation, (8.197) reduces to a Bessel equation of the form (8.210) with

$$\nu = \frac{3}{2} + \varepsilon. \quad (8.229)$$

The solution that satisfies the initial conditions (8.184) is given by (8.211). For super-Hubble modes, this solution should then be expanded, keeping the leading and next to leading order terms exactly in the same way as in the previous section. By noticing that

$$\mathcal{P}_T = 16\varepsilon \mathcal{P}_\zeta^{[\delta=\varepsilon]}, \quad (8.230)$$

we easily obtain

$$\mathcal{P}_T = \frac{16}{\pi} \frac{H^2}{M_p^2} [1 - 2(C+1)\varepsilon] \left(\frac{k}{aH} \right)^{-2\varepsilon}. \quad (8.231)$$

We then infer that the spectrum in the slow-roll regime takes the two equivalent forms,

$$\boxed{\mathcal{P}_T = \frac{16}{\pi} \frac{H^2}{M_p^2} \left[1 - 2(C+1)\varepsilon - 2\varepsilon \ln \left(\frac{k}{aH} \right) \right]}, \quad (8.232)$$

$$\boxed{= \frac{16}{\pi} [1 - 2(C+1)\varepsilon] \left(\frac{H^2}{M_p^2} \right)_{k=aH}.} \quad (8.233)$$

8.5.2.5 Consistency relation at leading order

The cosmological predictions of inflation are, in general, discussed in terms of the power spectra

$$A_s^2 \equiv \frac{4}{25} \mathcal{P}_\zeta(k), \quad A_T^2 \equiv \frac{1}{100} \mathcal{P}_T(k), \quad (8.234)$$

the normalization of these spectra being a pure matter of convention.

The spectral index of scalar perturbations appears in (8.227),

$$n_s - 1 = \frac{d \ln \mathcal{P}_\zeta}{d \ln k} = \frac{d \ln A_s^2}{d \ln k} = 2\delta - 4\varepsilon. \quad (8.235)$$

It is instructive to determine n_s from (8.228). Using (8.57), we obtain

$$\begin{aligned} n_s - 1 &= 2 \frac{d \ln H_k}{d \ln k} - \frac{d \ln \varepsilon_k}{d \ln k} = -2\varepsilon - \frac{d \ln \varepsilon_k}{dt} \times \frac{dt}{d \ln a_k} \frac{d \ln a_k}{d \ln k}, \\ &= -2\varepsilon - \frac{\dot{\varepsilon}}{\varepsilon} \times \frac{1}{H} \times 1 = 2\delta - 4\varepsilon. \end{aligned} \quad (8.236)$$

The spectral index of the tensor modes appears in (8.232)

$$n_T = \frac{d \ln \mathcal{P}_T}{d \ln k} = \frac{d \ln A_T^2}{d \ln k} = -2\varepsilon. \quad (8.237)$$

The index n_T can also be determined from (8.233). Using (8.57), we obtain

$$n_T = 2 \frac{d \ln H_k}{d \ln k} = -2\varepsilon. \quad (8.238)$$

These results summarize one of the generic predictions of inflation. In the slow-roll regime, inflation predicts that density perturbations and gravitational waves develop super-Hubble correlations with nearly scale invariant power spectra. This is due to the fact that the observable frequency band corresponds to a small number of e-folds and thus to small variations of φ .

Each spectrum is characterized by its amplitude and its spectral index. Even though the absolute amplitude is arbitrary and is fixed by the energy scale of inflation, H , the relative amplitudes of the scalar and tensor modes must satisfy the relation

$$R \equiv \frac{A_T^2}{A_s^2} = \varepsilon. \quad (8.239)$$

This implies a consistency relation between the amplitudes of the two spectra and their spectral indices,

$$R = -\frac{n_T}{2}. \quad (8.240)$$

This relation is incontrovertible and must be satisfied to lowest order by any model of single-field inflation in the slow-roll regime. It is a distinctive signature of inflation and it is difficult to imagine another mechanism that would lead to such an equality. Needless to say that checking this relation is a priority task for cosmology.

8.5.2.6 *Consistency relation at next to leading order*

At next order, the k dependence of the slow-roll parameters should be taken into account. Using (8.56) and (8.57), we obtain from the logarithmic derivatives of (8.228) and (8.233)

$$n_s - 1 = 2\delta - 4\varepsilon - [8(C+1)\varepsilon^2 - (6+10C)\varepsilon\delta + 2C\xi_H^2], \quad (8.241)$$

$$n_T = -2\varepsilon [1 + (3+2C)\varepsilon - 2(1+C)\delta]. \quad (8.242)$$

By replacing H' by ε in the ratio between A_s and A_T we find that

$$\varepsilon = R[1 - 2C(\varepsilon - \delta)], \quad (8.243)$$

which corresponds to (8.239) at next order. Moreover, from the ratio between (8.228) and (8.233) we find that

$$n_T = -2R(1 + 3\varepsilon - 2\delta). \quad (8.244)$$

After replacing ε and δ in the brackets by their values at leading order, we obtain the consistency relation at next to leading order

$$n_T = -2R[1 - R^2 + (1 - n_s)]. \quad (8.245)$$

It is remarkable that this expression does not involve any derivatives of the spectral indices. Moreover, since the scalar modes are easier to measure than the gravitational waves, it is probable that if n_T is known then so is n_s , so that if the consistency relation can be observationally verified at leading order, one can also verify it at next order without any additional effort.

As shown by relations (8.241) and (8.242), at this order n_s and n_T depend on the wave number k . We can therefore define

$$\alpha_s \equiv \frac{dn_s}{d \ln k}, \quad \alpha_T \equiv \frac{dn_T}{d \ln k}, \quad (8.246)$$

which characterize the running of the spectral indices.

Using (8.46) we find that

$$\frac{d\varepsilon}{d \ln k} = 2\varepsilon(\delta - \varepsilon), \quad \frac{d\delta}{d \ln k} = 2\varepsilon(\varepsilon - \delta) - \xi.$$

At lowest order, α_s and α_T are obtained by considering the value of the spectral indices at leading order so that

$$\alpha_s = -4\varepsilon(\varepsilon - \delta) - 2\xi, \quad \alpha_T = -4\varepsilon(\varepsilon - \delta). \quad (8.247)$$

This implies the existence of a second consistency relation [27]

$$\alpha_T = 2R[2R + (n_s - 1)]. \quad (8.248)$$

The measurement of α_T remains for now far from any observational possibilities, and so within the present state of technical possibilities, this relation is not yet very interesting.

Table 8.1 summarizes the results obtained for the spectral indices of inflationary spectra in the slow-roll regime. Note also that if these parameters are measured at a given so-called *pivot* scale, k_* , then the expansion of these spectra takes the form

$$\ln A_s^2(k) = \ln A_s^2(k_*) + [n_s(k_*) - 1] \ln \frac{k}{k_*} + \frac{1}{2} \alpha_s \ln^2 \frac{k}{k_*} + \dots, \quad (8.249)$$

and an analogous version for A_T^2 . k_* is usually chosen as the logarithmic mean of the observable frequency band, typically $k_* \sim 0.01 h \text{ Mpc}^{-1}$.

Table 8.1 Summary of the properties of the scalar modes and gravitational waves power spectra.

| | Leading order | Next to leading order |
|----------------------|--------------------------|---|
| R | ε | |
| $n_s - 1$ | $2\delta - 4\varepsilon$ | $2\delta - 4\varepsilon - [8(C+1)\varepsilon^2 - (6+10C)\varepsilon\delta + 2C\xi_H^2]$ |
| n_T | $-2H, \varepsilon$ | $-2\varepsilon[1 + (3+2C)\varepsilon - 2(1+C)\delta]$ |
| consistency relation | $n_T = -2R$ | $n_T = -2R[1 - R^2 + (1 - n_s)]$ |
| α_s | 0 | $-4\varepsilon(\varepsilon - \delta) - 2\xi$ |
| α_T | 0 | $-4\varepsilon(\varepsilon - \delta)$ |
| consistency relation | | $\alpha_T = 2R[2R + (n_s - 1)]$ |

8.5.3 Relation to observations

As long as we work in the slow-roll regime, perturbations are characterized by the two primordial spectra A_T and A_s . We briefly review what observables can teach us about them.

For super-Hubble modes, ζ is constant during the evolution. Computations of the matter power spectrum and temperature anisotropies of the cosmic microwave background require the knowledge of the spectrum of Φ . Equation (5.149) allows us to obtain

$$\Phi(k\eta \ll 1, \eta < \eta_{\text{eq}}) = \frac{2}{3}\zeta, \quad \Phi(k\eta \ll 1, \eta > \eta_{\text{eq}}) = \frac{3}{5}\zeta. \quad (8.250)$$

8.5.3.1 Cosmic microwave background

The previous results fix the initial conditions for the computation of the angular power spectrum of the cosmic microwave background anisotropies (Chapter 6). Restricting ourselves to large angular scales, i.e. to small multipoles, the angular spectrum of the scalar modes is given by (6.50), using $\delta T/T \sim \frac{1}{3}\Phi$ for adiabatic perturbations at long wavelengths. Since $\Phi = \frac{3}{5}\zeta$, we infer that

$$C_\ell^S \simeq \pi A_s^2(k_0) \left[\frac{\pi}{2} \frac{\Gamma(3 - n_s)}{2^{3-n_s}} \frac{\Gamma\left(\ell + \frac{n_s - 1}{2}\right)}{\Gamma^2\left(2 - \frac{n_s}{2}\right) \Gamma\left(\ell + \frac{5 - n_s}{2}\right)} \right], \quad (8.251)$$

if we choose the pivot $k_* = k_0 = 1/\eta_0$. For $n_s \sim 1$, we obtain $\ell(\ell+1)C_\ell^S = \pi A_s^2(k_0)/2$. Observations of the anisotropies by the satellite WMAP [28] impose that

$$Q_{\text{rms-PS}} \equiv \Theta_0 \sqrt{\frac{5C_2^S}{4\pi}} = 18 \times 10^{-6} \text{ K}. \quad (8.252)$$

For $\Theta_0 = 2.73$ K, we infer that $A_s^2(k_0) = 4.16 \times 10^{-10}$. Using (8.213) and (8.49) we infer the energy scale of inflation

$$\left(\frac{V}{\varepsilon}\right)^{1/4} \simeq 0.0047 M_p \simeq 1.5 \times 10^{17} \text{ GeV}. \quad (8.253)$$

This value is actually an upper bound as part of the perturbations can be of tensorial origin.

Performing the same computation for gravitational waves, the consistency relation becomes

$$r = \frac{C_2^S}{C_2^T} = f(h, \Omega, \Omega_\Lambda, \dots) \varepsilon. \quad (8.254)$$

The function f depends on a given cosmological model. For the flat standard model with $\Omega_{\Lambda 0} = 0.7$, $f \sim 10$.

Analysis from the WMAP data provides the constraint [30] on ε

$$\varepsilon < 0.032, \quad (8.255)$$

which implies

$$\left(\frac{H_*}{M_p}\right) \lesssim 1.4 \times 10^{-5} \iff \left(\frac{V_*}{M_p^4}\right)^{1/4} \lesssim 2.2 \times 10^{-3}. \quad (8.256)$$

The reanalysis of the WMAP data [29] leads to the constraints that $R < 0.0267$ while $16R < 0.0125$ when combined with baryon acoustic oscillations and supernovae data for a tensor-to-scalar ratio at $k = 0.002 \text{ Mpc}^{-1}$. The scalar spectral index is, respectively, obtained to be $n_s = 0.963_{-0.015}^{+0.014}$ and $n_s = 0.960_{-0.013}^{+0.014}$ at a 68% confidence level.

8.5.3.2 Link with the number of e-folds

Observations allow us to establish constraints on the pair $(r, n_s - 1)$ that can be related to the slow-roll parameter space. These quantities depend on the form of the potential but also on the number of e-folds, N , between the time when the largest observable scales become super-Hubble and the end of inflation.

To illustrate this, let us consider a model of inflation with large field values (model of type A), for instance, with a power law potential (8.68). The expressions (8.69) and (8.235) allow us to establish that

$$r = \frac{f}{2} \left(\frac{n}{n+2} \right) (1 - n_s). \quad (8.257)$$

Inverting (8.70) and (8.69) for ε , we obtain

$$r = f \frac{n}{4(N+n/4)}, \quad 1 - n_s = \frac{n+2}{2(N+n/4)}. \quad (8.258)$$

For a given model, for example, $n = 4$, the observable quantities r and $n_s - 1$ depend on the number of e-folds. Figure 8.11 illustrates the position of one of these models in the $(n_s - 1, r)$ -plane. As N is increased the model gets closer and closer to the point $(n_s - 1, r) = (0, 0)$. This figure also summarizes the constraints on the two parameters obtained from WMAP data.

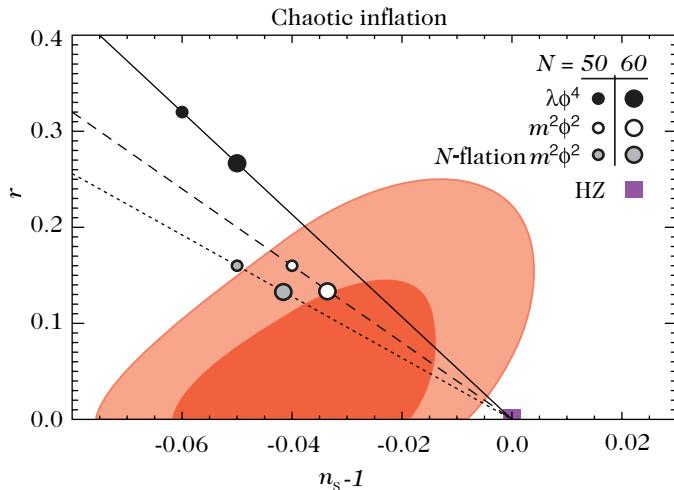


Fig. 8.11 Constraints on single-field chaotic inflationary models with potentials of the form φ^n with $n = 2$ (dashed), $n = 4$ (solid) and on Nflation with φ^2 potential (dotted). HZ is the prediction for a strictly scale invariant power spectrum. The predictions for $N = 50$ and $N = 60$ e-folds have been plotted. The φ^4 models are excluded at 95% CL. From Ref. [29].

Given observational constraints in this plane, the viability of a model depends on the number of e-folds. The previous example shows that predictions from a model of chaotic inflation depend crucially on N . The latest analysis of WMAP [29] concluded that a single-field model with $V = \lambda\varphi^4/4$ is far from the 95% confidence level (CL) region for both $N = 50$ and $N = 60$. A massive free-field model is out of the 68% CL region for $N = 50$ and at the boundary of this region for $N = 60$ while being inside the 95% CL region. For a power-law inflation model, $R = 1/p$ and $1 - n_s = 2/p$ so that $p < 60$ is excluded at more than 99% CL.

The symmetry with respect to the end of inflation (for $\eta < \eta_e$, the modes exit the Hubble radius and for $\eta > \eta_e$, the modes enter the Hubble radius, see Fig. 8.5) allows us to put an upper bound on N from H_e and H_0 . One can show [31] that

$$N < 60.9 + \frac{1}{2} \ln \left(\frac{H_*}{10^{15} \text{ GeV}} \right) - \ln \left(\frac{k}{0.02 \text{ Mpc}^{-1}} \right). \quad (8.259)$$

Using $H_*/M_p \lesssim 2 \times 10^{-5}$ and that the largest observable scale is $k = a_0 H_0$ we can then estimate that

$$N < 62.5 + \ln \left(\frac{0.6}{h} \right). \quad (8.260)$$

A precise determination of N is important for a good understanding of the end of inflation. An upper bound for the energy scale of inflation can be obtained by assuming that the part of the angular power spectrum that corresponds to the Sachs–Wolfe plateau is due to gravitational waves only. We then obtain

$$V_*^{1/4} < 3 \times 10^{16} \text{ GeV}.$$

8.5.3.3 Consistency relation

In order to check the consistency relation, the detection of primordial gravitational waves is necessary. This is beyond current observations. We summarize here the constraints obtained on the contributions of the gravitational waves.

Analysis from WMAP data has, for example, established [30] that

$$\frac{C_{10}^T}{C_{10}^S} < 0.32,$$

so that gravitational waves do not contribute more than 30% of the cosmic microwave background.

Relation (8.42) allows us to establish that during an interval of ΔN number of e-folds, the field has moved by

$$\frac{\Delta\varphi}{M_p} = \sqrt{\frac{\varepsilon}{4\pi}} \Delta N, \quad (8.261)$$

which, using (8.254), can be re-expressed as

$$\frac{\Delta\varphi}{M_p} = \sqrt{\frac{16\pi r}{f}} \Delta N \sim 2.37 \left(\frac{r}{0.07} \right)^{1/2} \left(\frac{f}{10} \right)^{-1/2} \left(\frac{\Delta N}{4} \right). \quad (8.262)$$

This estimate allows us to conclude [32] that if $r \sim 0.1$, then the field must have varied by more than M_p . This result has often been used to argue that realistic potentials for inflation imply that r is very small.

Due to cosmic variance, the contribution of gravitational waves to temperature anisotropies can only be distinguished if $r > 0.07$ [33]. Primordial gravitational waves generate anisotropies in the B polarization (Chapter 6). These B modes are also

generated from the E polarization by gravitational lensing by the large-scale structures. This foreground effect limits the detection of primordial gravitational waves to $r > 6 \times 10^{-4}$ [34], which implies that the energy scale of inflation has to be larger than

$$V_*^{1/4} > 3.2 \times 10^{15} \text{ GeV} \iff \left(\frac{V_*}{M_{\text{P}}^4} \right)^{1/4} \gtrsim 2.62 \times 10^{-4}. \quad (8.263)$$

This is only one order of magnitude smaller than the current limit (8.256). This limit can be lowered by a factor of 2 if the variation of the spectral index is measured. These considerations show that the consistency relation will be difficult to verify observationally.

8.5.4 Reconstruction of the potential

8.5.4.1 Principle

The measurement of n_s , n_t , R , α_s and α_t must, in principle, make it possible to reconstruct the form of the potential in the neighbourhood of φ_* , the value of φ at the moment where the pivot becomes super-Hubble. To see this, one should establish the link between the expansion (8.249) and a Taylor expansion of the potential around φ_*

$$V = V_* + V'_* \Delta\varphi + \frac{1}{2} V''_* (\Delta\varphi)^2 + \dots \quad (8.264)$$

The definitions of the slow-roll parameters can be used to express the successive derivatives of the potential as

$$V_* = \frac{M_{\text{P}}^2}{8\pi} H^2 (3 - \varepsilon), \quad (8.265)$$

$$V'_* = -\frac{M_{\text{P}}}{\sqrt{4\pi}} H^2 \sqrt{\varepsilon} (3 - \delta), \quad (8.266)$$

$$V''_* = H^2 (3\varepsilon + 3\delta - \delta^2 - \xi_H^2). \quad (8.267)$$

As a second step, we express the slow-roll parameters as a function of observables, such as the spectral index. For instance, using expressions (8.233) and (8.243), we find that V_* is given by

$$V_* = \frac{75}{32} M_{\text{P}}^4 A_t^2(k_*) \left[1 + \left(\frac{5}{3} + 2C \right) R_* \right]. \quad (8.268)$$

The detail of the computation of the other derivatives can be obtained in Ref. [35]. Here, we give just the relations for the two following derivatives without any demonstration

$$\begin{aligned} V'_* &= -\frac{75}{8} \sqrt{\pi} M_{\text{P}}^3 R_*^{3/2} A_s^2(k_*) \left[1 + \left(3C + \frac{4}{3} \right) R_* - \frac{1}{2} \left(C - \frac{1}{3} \right) (1 - n_{S*}) \right], \\ V''_* &= \frac{25}{4} \sqrt{\pi} M_{\text{P}}^2 R_* A_s^2(k_*) \left[9R_* - \frac{3}{2}(1 - n_{S*}) + (36C + 2)R_*^2 \right. \\ &\quad \left. - \frac{1}{4}(1 - n_{S*})^2 - (12C - 6)R_*(1 - n_{S*}) - \frac{3C - 1}{2}\alpha_{S*} \right]. \end{aligned} \quad (8.269)$$

Table 8.2 summarizes the different observables that are necessary to reconstruct the potential up to a given order. Note once again that the absolute normalization of the potential requires knowledge of A_T , i.e. the detection of primordial gravitational waves.

Table 8.2 Summary of the slow-roll parameters and observables that are necessary to reconstruct a derivative of given order of the potential, at leading and next order.

| | Leading order | Next order | Leading order | Next order |
|--------|-------------------------------|-------------------------------|---------------------------|---------------------------|
| V | H | H, ε | A_T^2 | A_T^2, R |
| V' | H, ε | H, ε, δ | A_T^2, R | A_T^2, R, n_S |
| V'' | H, ε, δ | $H, \varepsilon, \delta, \xi$ | A_T^2, R, n_S | A_T^2, R, n_S, α_S |
| V''' | $H, \varepsilon, \delta, \xi$ | | A_T^2, R, n_S, α_S | |

8.5.4.2 Flow equations

Another way to tackle the problem is to randomly generate potentials and to deduce the observables [36]. For this, we extend the hierarchy of the slow-roll parameters to all orders by defining the quantities

$${}^\ell\lambda_H \equiv \left(\frac{M_P^2}{4\pi} \right)^\ell \frac{(H')^{\ell-1}}{H^\ell} \frac{d^{(\ell+1)}H}{d\varphi^{(\ell+1)}}, \quad (8.270)$$

which complete the doublet $(\varepsilon_H, \delta_H)$. In particular, $\delta_H = {}^1\lambda_H$ and $\xi_H^2 = {}^2\lambda_H$. Using as a variable the number of e-folds before the end of inflation, we have

$$\frac{d}{dN} = \frac{d}{d \ln a} = \frac{M_P}{\sqrt{4\pi}} \sqrt{\varepsilon} \frac{d}{d\varphi}.$$

We have chosen the sign of $\sqrt{\varepsilon}$ to be the same as that of H' . Having $dt > 0$ thus corresponds to $d\varphi < 0$ and $dN < 0$. The evolution equations (8.46) can then be extended into the series of equations

$$\begin{aligned} \frac{d\varepsilon_H}{dN} &= 2\varepsilon_H(\delta_H - \varepsilon_H), \\ \frac{d^\ell\lambda_H}{dN} &= [(\ell-1)^1\lambda_H - \ell\varepsilon_H] {}^\ell\lambda_H + {}^{\ell+1}\lambda_H. \end{aligned} \quad (8.271)$$

This system can be integrated numerically to an arbitrary order in the slow-roll parameters. In any practical application, this hierarchy is truncated to a given value $\ell = M$. We then choose a number of e-folds, N , and a random point in the space $\{\varepsilon, {}^\ell\lambda_H\}$ and evolve (8.271) until the end of inflation or until a fixed point of the system is reached. In the second case, we compute the observables (r, n_S, \dots) . In the first case, the system is integrated backwards in time for N e-folds and observables are computed at this point. Figure 8.12 illustrates the distribution of the integration of 10^6 models. In particular, 90% of the potentials have a blue-tilt spectrum with $n_S > 1$ and are thus excluded.

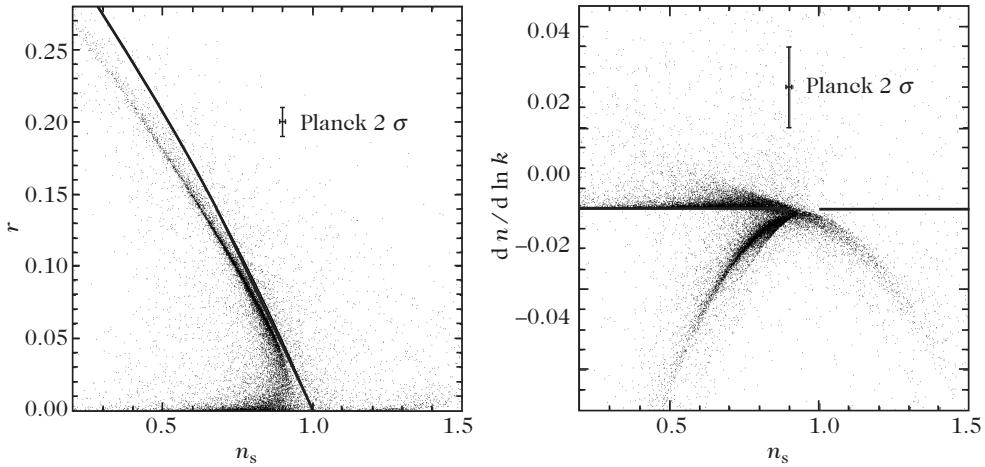


Fig. 8.12 Distribution of 10^6 models with $M = 5$ and $N = 40-70$ in the space of observables. From Ref. [36].

If we have observational constraints in the space of observables, we can reconstruct the potential if the point is within acceptable limits. This method can be used to explore the space of potentials and to extract those that agree with observations. However, nothing ensures that the reconstructed potentials are realistic from a theoretical point of view and it is difficult to go beyond a simple catalogue of potentials since the measure, in the sense of probability measure, in the space of potentials is unknown, which makes it impossible to put relative weights on the models. This method has been used to compare theory with observations [36] but the link between the flow equations and the dynamics of inflation is far from clear [37].

8.6 End of inflation and reheating

The theory of reheating after inflation is one of the most important applications of quantum field theory in curved space-time. Almost all the matter present in our Universe would have been created through a process of particle production during the inflaton decay.

8.6.1 Perturbative reheating

During inflation, all the energy is concentrated in the slowly rolling inflaton. Shortly after the end of inflation, the Universe is cold and ‘frozen’ in a state of low entropy where the field oscillates around the minimum of its potential (Fig. 8.6). The coherent oscillations of the inflaton can be considered as a collection of independent scalar particles. If they couple to other particles, the inflaton can then decay perturbatively to produce light particles. The interaction of the inflaton should therefore give rise to an effective decay rate, Γ_φ , and reheating would only occur after the expansion rate had decreased to a value $H \sim \Gamma_\varphi$. This also implies that during the first $\mathcal{O}(m/\Gamma_\varphi)$ oscillations of the inflaton, nothing interesting happens.

Let us illustrate this mechanism in the case of a model of chaotic inflation with the potential (8.68). As illustrated in Fig. 8.6 for $n = 2$, the inflaton behaves as a pressureless fluid during the oscillatory phase and the scale factor evolves as $a \propto t^{2/3}$. To see this, let us rewrite the Klein–Gordon equation as

$$\frac{d}{dt} (a^3 \rho_\varphi) = -P_\varphi \frac{da^3}{dt}.$$

During the oscillatory phase, $H < m$ and the inflaton undergoes several oscillations during a time H^{-1} . It is thus reasonable to use the mean value of the pressure and density over several oscillations. For a harmonic oscillator $\langle \dot{\varphi}^2/2 \rangle = (n/2)\langle V(\varphi) \rangle$, so that $\langle P_\varphi \rangle = (n-2)/(n+2)\langle \rho_\varphi \rangle$. The scalar field thus behaves as a dust fluid for $n = 2$ and as a radiation fluid for $n = 4$.

In order for the inflaton to decay, it should be coupled to other fields. Let us restrict ourselves to the case $n = 2$ and, as an example, let us consider couplings described by the Lagrangian

$$-\mathcal{L}_{\text{int}} = \sigma v \varphi \chi^2 + h \varphi \bar{\psi} \psi, \quad (8.272)$$

where σ and h are dimensionless coupling constants and v is a mass scale. No mass term has been introduced for the bosons χ and the fermions ψ that are assumed to be very light compared to the inflaton. This is necessary in order for the decay rate of the inflaton into these particles to be noticeable. As seen earlier, the mass scale of the inflaton should be of the order of $m \sim 10^{-6} M_P$.

8.6.1.1 Evolution of the inflaton

Including quantum corrections, the Klein–Gordon equation becomes

$$\ddot{\varphi} + 3H\dot{\varphi} + [m^2 + \Pi(m)] \varphi = 0, \quad (8.273)$$

where $\Pi(m)$ is the polarization operator of the inflaton. The real part of $\Pi(m)$ corresponds to the mass correction. Even though $\text{Im}[\Pi(m)] \ll m^2$, Π has an imaginary part

$$\text{Im}[\Pi(m)] = m \Gamma_\varphi. \quad (8.274)$$

Since m is much larger than both Γ_φ and H at the end of inflation, we can solve the Klein–Gordon equation by assuming that both these quantities are constant during an oscillation. We infer that the inflaton evolves as

$$\varphi = \Phi(t) \sin mt, \quad \Phi = \varphi_0 \exp \left[-\frac{1}{2} \int (3H + \Gamma_\varphi) dt \right]. \quad (8.275)$$

As long as $3H > \Gamma_\varphi$, the decrease in the inflaton energy caused by the expansion (Hubble friction) dominates over particle decay. Since $H = 2/(3t)$, we infer that

$$\Phi = \varphi_f \frac{t_f}{t} = \frac{M_P}{m} \frac{1}{\sqrt{3\pi t}} \quad (8.276)$$

where we have used $\varphi_f = M_P/\sqrt{4\pi}$, $t_f = 2/3H_f$ and $H_f^2 = (4\pi/3)(m/M_P)^2 \varphi_f^2$. We recover $\langle \rho_\varphi \rangle = m^2 \Phi^2/2 \propto a^{-3}$.

8.6.1.2 Reheating temperature

Reheating can only occur when $\Gamma_\varphi \gtrsim 3H$. In this regime, we find that

$$\Phi = \frac{M_p}{m} \frac{1}{\sqrt{3\pi t}} e^{-\Gamma_\varphi t/2}. \quad (8.277)$$

We define the time of reheating, t_{reh} , by $\Gamma_\varphi = 3H$ so that the energy density at that time is

$$\rho_{\text{reh}} = \frac{\Gamma_\varphi^2 M_p^2}{24\pi}. \quad (8.278)$$

If this energy is rapidly converted into radiation, its temperature is

$$\rho_{\text{reh}} = \frac{\pi^2}{30} g_* T_{\text{reh}}^4. \quad (8.279)$$

The reheating temperature is thus given by

$$T_{\text{reh}} = \left(\frac{5}{4\pi^3 g_*} \right)^{1/4} \sqrt{\Gamma_\varphi M_p} \simeq 0.14 \left(\frac{100}{g_*} \right)^{1/4} \sqrt{\Gamma_\varphi M_p} \ll 10^{15} \text{ GeV}, \quad (8.280)$$

where the upper bound was obtained from the constraint $\Gamma_\varphi \ll m \sim 10^{-6} M_p$, assuming $g_* \gtrsim 100$. Note that $T_{\text{reh}} < E_{\text{GUT}}$, the characteristic scale of Grand Unified symmetry breaking (see Chapter 9), so that baryogenesis cannot happen at this scale in such a reheating scenario. Moreover, in order for gravitinos not to be overproduced, $T_{\text{reh}} \lesssim 10^9 - 10^{10}$ GeV is required.

8.6.1.3 Decay rate

Γ_φ is the total decay rate of the inflaton and can be decomposed as $\Gamma_\varphi = \Gamma(\varphi \rightarrow \chi\chi) + \Gamma(\varphi \rightarrow \bar{\psi}\psi)$ where the decay rates into bosons and fermions are, respectively, given by

$$\Gamma(\varphi \rightarrow \chi\chi) = \frac{\sigma^2 v^2}{8\pi m}, \quad \Gamma(\varphi \rightarrow \bar{\psi}\psi) = \frac{h^2 m}{8\pi}. \quad (8.281)$$

These expressions are only valid in the limit where $\Gamma_\varphi \ll m$, i.e. $h \ll 1$ and $\sigma v \ll m$. Assuming $h \sim 10^{-2}$, the decay rate into fermions is of the order of $\Gamma(\varphi \rightarrow \bar{\psi}\psi) \sim 10^8$ GeV. The reheating temperature (8.280) is then $T_{\text{reh}} \sim 10^{13}$ GeV and the inflaton undergoes about 10^5 oscillations before decaying. For bosons, assuming $\sigma \sim 10^{-2}$ and $v \sim 10^{11}$ GeV, we find that $\Gamma(\varphi \rightarrow \chi\chi) \sim 10^3$ GeV only. The reheating temperature is of the order of $T_{\text{reh}} \sim 10^{10}$ GeV and the inflaton undergoes about 10^{10} oscillations. Note that the decay rate for an interaction of the type $\mathcal{L}_{\text{int}} = -g^2 \varphi^2 \chi^2 / 2$ is

$$\Gamma(\varphi\varphi \rightarrow \chi\chi) = \frac{g^2 \varphi^2}{8\pi m}. \quad (8.282)$$

Unlike ‘three-leg’ interactions, this rate depends on the value of the inflaton. Since $\varphi^2 \sim a^{-3} \sim t^{-2}$, this decay rate decreases quicker than the Hubble rate H . Perturbative reheating can therefore never be produced by ‘four-leg’ interactions.

Complete reheating can only occur if Γ_φ decreases slower than t^{-1} . Hence assuming that reheating is complete imposes constraints on the structure of the theory and on the couplings of the inflaton with other fields.

510 Inflation

8.6.1.4 Phenomenological formulation

To finish, let us mention a simplified version of this mechanism in which we introduce the coupling between the scalar field and radiation. The conservation equations are then transformed into

$$\dot{\rho}_\varphi + H(\rho_\varphi + P_\varphi) = -\Gamma_\varphi \rho_\varphi, \quad \dot{\rho}_r + 4H\rho_r = \Gamma_\varphi \rho_\varphi. \quad (8.283)$$

Reheating can then be described phenomenologically as a transfer between two fluids.

8.6.2 Theory of preheating

The theory of perturbative reheating is simple and intuitive in many aspects. However, the decay of the inflaton can start much earlier in a phase of *preheating* (parametric reheating) where particles are produced by parametric resonance.³ The preheating process can be decomposed into three stages: (1) non-perturbative production of particles, (2) perturbative stage and (3) thermalization of the produced particles.

8.6.2.1 Parametric resonance

In order to illustrate this mechanism, let us consider a model of chaotic inflation with potential (8.62) and an interaction of the form

$$V_{\text{int}} = \frac{1}{2}g^2\varphi^2\chi^2. \quad (8.284)$$

The scalar field χ can be decomposed, in second quantization, into

$$\hat{\chi} = \int \frac{d^3k}{(2\pi)^{3/2}} [\chi_k(t)\hat{a}_k(t)e^{-ik\cdot x} + \text{h.c.}] . \quad (8.285)$$

Considering only the interaction between this quantum operator and the classical field φ , then the evolution of each mode is dictated by

$$\ddot{\chi}_k + 3H\dot{\chi}_k + \left(\frac{k^2}{a^2} + g^2\varphi^2\right)\chi_k = 0. \quad (8.286)$$

One can rewrite this equation in terms of the variable $X_k = a^{3/2}\chi_k$ as

$$\ddot{X}_k + \omega_k^2 X_k = 0, \quad \omega_k^2 = \frac{k^2}{a^2} + g^2\varphi^2 - \frac{3}{4}(2\dot{H} + 3H^2). \quad (8.287)$$

Using the time variable $z = mt$ and defining the quantities

$$q \equiv \frac{g^2\Phi^2(t)}{4m^2}, \quad A_k \equiv 2q + \frac{k^2}{m^2a^2}, \quad (8.288)$$

where Φ is defined by (8.275), this equation takes the form

³Defining a parametric oscillator as an oscillator with time-dependent frequency, the associated resonance, when it exists, will be a parametric resonance.

$$\frac{d^2}{dz^2}X_k + (A_k - 2q \cos 2z + \Delta)X_k = 0, \quad (8.289)$$

where $\Delta = [m_\chi^2 - 3(2\dot{H} + 3H^2)/4]/m^2$. This term is, in general, negligible, since $m_\chi/m \ll 1$ and, e.g., if $H = 2/3t$, so that $2\dot{H} + 3H^2 = 0$. In fact, for $\Delta \ll 1$ and A_k and q constant, this equation is a *Mathieu equation*.

8.6.2.2 Minkowski space-time case

For a Minkowski space-time, the coefficients A_k and q are constant (we set $a = 1$). Then this equation always admits a solution of the form $X_k(z) = P_k(z) \exp(i\nu_k z)$, where P_k is a periodic function of period π . ν_k is the characteristic exponent; it is not defined in a unique way as the solution is invariant under $\nu_k \rightarrow \nu_k + 2n$. The solution is stable when $\text{Im}(\nu_k) > 0$. In the opposite case, this equation has instabilities and the solution then increases exponentially as $X_k \propto \exp(\mu_k z)$ with $\nu_k = -i\mu_k$ with $\mu_k > 0$. The properties of the solutions of this equation are then described by a stability diagram (Fig. 8.13). Note that the Mathieu equation is simply the Schrödinger equation for an electron in a periodic potential. In this case the unstable zones correspond to forbidden bands in the energy spectrum.

As illustrated in Fig. 8.13, we obtain frequency bands $\Delta_n k$ to which we can associate an integer index n . For a mode $k \in \Delta_n k$, the instability corresponds to an exponential growth in the occupation number $n_k \propto \exp(2\mu_k^{(n)} z)$ that can be interpreted as a production of particles. In this regime, called *preheating*, the perturbative description of Γ_φ is not applicable.

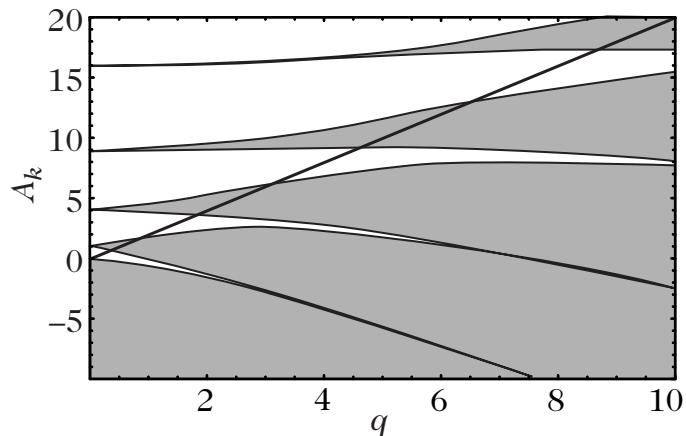


Fig. 8.13 Stability diagram of the Mathieu equation. The shaded zones are instability zones where X_k increases in an exponential way. The diagonal line represents $A_k = 2q$.

8.6.2.3 Resonances

When $q \lesssim 1$, the width of the instability band is narrow. In this regime and for the first band ($n = 1$), it corresponds to modes $k^2 \sim m^2(1 - 2q \pm q)$. The central mode of the resonance grows as $\exp(qz/2)$, i.e. as $\exp(q^2\Phi^2t/8m)$. The number of particles, defined as

$$n_k = \frac{\omega_k}{2} \left(\frac{|\dot{\chi}_k|^2}{\omega_k^2} + |\chi_k|^2 \right) - \frac{1}{2}, \quad (8.290)$$

then grows as $\exp(2\mu_k z) \sim \exp(qz)$. This process can be interpreted as the resonant decay of two particles φ of mass m into two particles χ with momentum $k \sim m$.

When Φ is large, q can also be very large and the resonance exists for a large band of modes. The index μ_k can then be very large and the reheating becomes very efficient. The resonance occurs for the modes $k^2/m^2 = A_k - 2q$, i.e. above the line $A_k = 2q$ of Fig. 8.13. Unlike for narrow resonances, the field χ_k oscillates many times, around $\mathcal{O}(\sqrt{q})$ times, during one oscillation of φ . During the main part of an oscillation, the effective mass of χ_k , $m_\chi = g\varphi(t)$, is larger than m so that the frequency of the oscillations of χ_k is typically $\omega_k(t) \simeq \sqrt{k^2 + g^2\varphi^2(t)}$. In this adiabatic regime, the production of particles is weak and n_k remains more or less constant. Production of particles requires that the variation of ω_k is no longer adiabatic

$$\dot{\omega}_k \gtrsim \omega_k^2, \quad (8.291)$$

which happens in the neighbourhood of $\varphi = 0$. For small values of φ , one has $\dot{\varphi} \sim m\Phi$, so that this condition translates into $k^2 \lesssim (g^2 m \varphi \Phi)^{2/3} - g^2 \varphi^2$. It can only be satisfied when $\varphi < 2\varphi_* = \sqrt{m\Phi/g}$. The maximal interval in k for which there is production of particles corresponds to $\varphi \lesssim \varphi_* \sim \Phi q^{-1/4}/3$, for which the maximal wave number is $k_{\max} \sim \sqrt{gm\Phi/2}$. Thus, if both q and the oscillation amplitude are large, particles with large momentum can be produced. The estimate of the behaviour of the parametric resonance given here is to be compared with the numerical solution of Fig. 8.14. One should note that the behaviour of n_k is much more regular and physically transparent than that of χ_k .

8.6.2.4 Expanding space-time

In an expanding space-time, the resonances only appear for $q^2 m \gtrsim H$, i.e. when

$$g\Phi \gtrsim 2m \left(\frac{H}{m} \right)^{1/4}. \quad (8.292)$$

At the end of inflation $H \sim m$ so that the explosive production of particles can only happen if $\Phi > m/g$, i.e. $q \gtrsim 1$. Hence, narrow resonances only play a role in the late phases of preheating but in this regime the backreaction induced by the particles χ_k has to be taken into account. This complicates its description.

The expansion modifies the picture obtained in Minkowski space-time by (1) red-shifting the momenta so a given mode will cross different instability zones of the Mathieu equation, (2) damping the amplitude of the oscillations and thus of q , which varies as $q \propto t^{-2}$. Thus, the frequency of the oscillations of the field χ also decreases

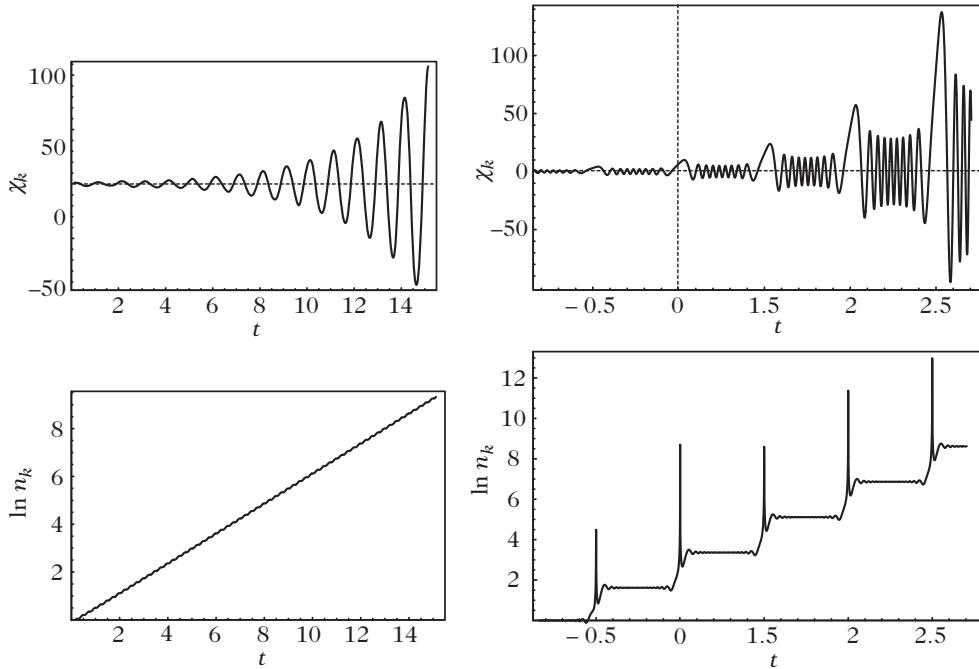


Fig. 8.14 Evolution of χ_k and of the number of particles n_k as a function of time in units of $m/2\pi$ in Minkowski space-time. (left): Narrow parametric resonance with $q \sim 0.1$. The number of particles grows in an exponential way with slope $\mu_k \sim q/2 \sim 0.05$. (right): Large resonance with $q \sim 200$. At each oscillation of φ , χ_k oscillates several times. The peaks correspond to the moments where $\varphi = 0$, where n_k is not defined. Adapted from Ref. [38].

with time. In particular, this changes the phase of χ_k at the time when $\varphi = 0$ in an unpredictable way so that the number of particles can either increase or decrease. This is what is called the phase of *stochastic resonance*. One can show that n_k increases statistically three times more often than it decreases so that in the end n_k still increases exponentially.

If $q \gg 1$ initially, there is a phase of stochastic resonance associated with large resonances. As q decreases, the preheating continues through narrow resonances and then stops. This succession of events is illustrated in Fig. 8.15. The complete theory describing this scenario is developed in Ref. [38].

8.6.2.5 Backreaction

The dynamics presented above is, however, too simplified because the field χ has been treated as a test field evolving in a space-time described by $a(t)$ and $\varphi(t)$ that are not themselves affected by the particle production. In fact, due to the exponential growth of χ_k , this field must backreact on the space-time dynamics and on the evolution of the field φ , in particular via a contribution of order $g^2 \langle \chi^2 \rangle$ to the effective mass of φ .

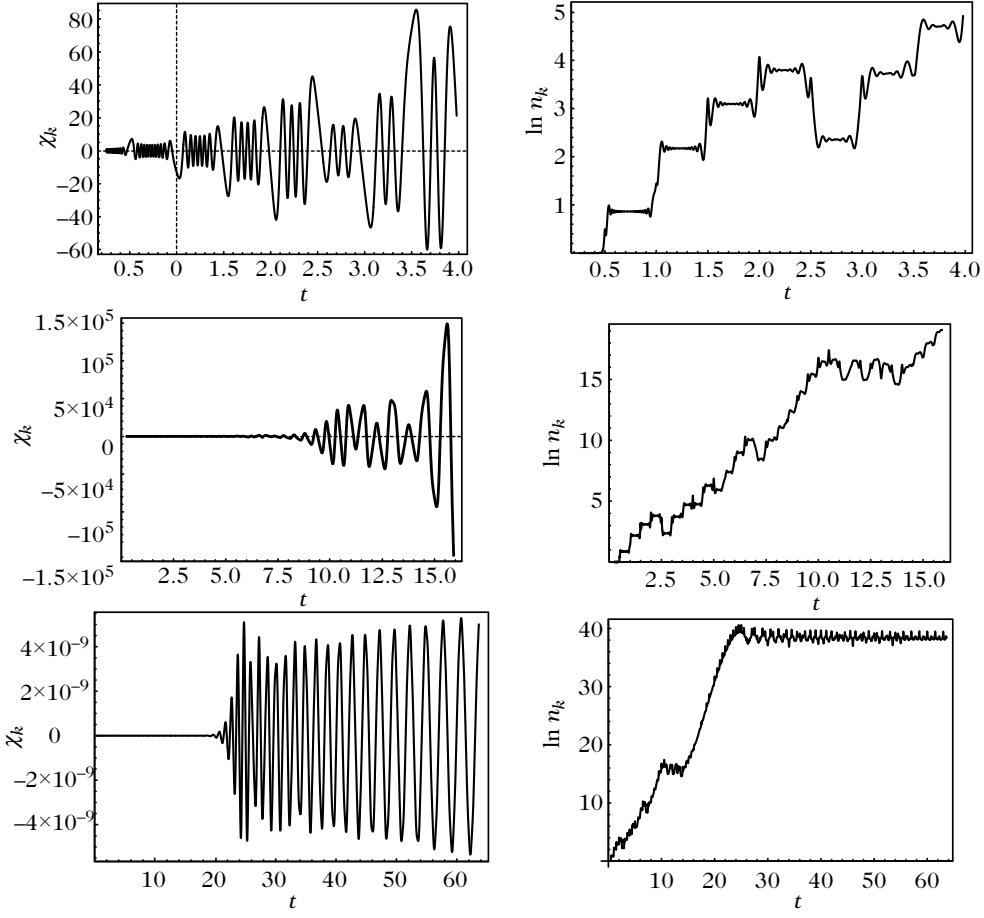


Fig. 8.15 Evolution of χ_k and of the number of particles n_k as a function of time in units of $m/2\pi$ (i.e. number of oscillations) in an expanding space-time for $q = 3 \times 10^3$ initially ($g = 5 \times 10^{-4}$, $m = 10^{-6} M_P$). During the first oscillations (top), n_k increases but can occasionally decrease, a purely quantum effect that would be impossible if the particles were created in a state of thermodynamic equilibrium. As t increases, $q \propto t^{-2}$ and the resonance becomes narrow (middle), then it disappears and n_k becomes constant (bottom). Adapted from Ref. [38].

A more precise description requires, first, to take into account the contribution of χ to the evolution of the Universe,

$$H^2 = \frac{8\pi G_N}{3} \left(\frac{1}{2}\dot{\varphi}^2 + \frac{1}{2}\dot{\chi}^2 + \frac{1}{2}m^2\varphi^2 + \rho_\chi + \rho_\varphi \right), \quad (8.293)$$

where ρ_χ and ρ_φ are the energy densities of the particles (i.e. $\mathbf{k} \neq 0$) of χ and φ .

We should then describe the coupled evolution of the perturbations of χ and φ , which introduces the polarization tensor for each field,

$$\ddot{X}_k + \left(\frac{k^2}{a^2} + g^2 \Phi^2 \sin^2 mt \right) X_k = - \int X_k(t') \Pi_\chi(t, t'; \mathbf{k}) dt', \quad (8.294)$$

$$\ddot{v}_k + \left(\frac{k^2}{a^2} + g^2 \Phi^2 \sin^2 mt \right) v_k = - \int v_k(t') \Pi_\varphi(t, t'; \mathbf{k}) dt', \quad (8.295)$$

where $v = a^{3/2} \delta\varphi$. To go further, we should therefore compute the polarization tensors. The description of the preheating phase then becomes more complicated and requires numerical simulations to be performed.

A simplified way to address the question is to work in the Hartree approximation for which the backreaction of χ on the evolution of the homogeneous solution is taken into account as

$$\ddot{\varphi} + 3H\dot{\varphi} + (m^2 + g^2 \langle \chi^2 \rangle) \varphi = 0, \quad (8.296)$$

with

$$\langle \chi^2 \rangle = \int \frac{k^2 dk}{2\pi^2 a^3} |X_k(t)|^2. \quad (8.297)$$

This term tends to make the oscillations of the inflaton incoherent and to suppress the resonance. Typically, the inflaton undergoes a few dozen oscillations and produces a number of particles of the order of $n_k \sim 10^2 g^{-2}$.

8.6.2.6 Discussion

The results of preheating are very model dependent. For instance, with the potential

$$V = \frac{\lambda}{4} \varphi^4 + \frac{1}{2} g^2 \varphi^2 \chi^2, \quad (8.298)$$

the resonance never becomes stochastic as all the parameters of the resonance have the same behaviour in Φ . The most general case with a term in $m^2 \varphi^2$ and a term in $\lambda \varphi^4$ gives rise to various possibilities depending on the value of the parameters [38].

Another important case is that of tachyonic reheating [39] that relies on an effective potential of the form $V = \lambda(\varphi^2 - v^2)^2/4 = \lambda\varphi^4/4 - m^2\varphi^2/2 + \dots$. The beginning of the phase transition ($\varphi \sim 0$) is associated with a tachyonic instability ($V'' < 0$). Any mode $k < m$ grows exponentially and

$$\langle \delta\varphi^2 \rangle = \int_0^m \frac{k dk}{4\pi^2} e^{2t\sqrt{m^2-k^2}},$$

until $\langle \delta\varphi^2 \rangle \sim v^2/4$. This happens at around $t_* \sim (2m)^{-1} \ln(\pi^2/\lambda)$ and the occupation number grows until

$$n_k \sim \exp(2mt_*) \sim \frac{\pi^2}{\lambda} \gg 1.$$

In this case, the field decays before reaching the minimum of the potential and the reheating phase is quasi-instantaneous.

The dynamics of preheating is therefore very different from the picture proposed by perturbative reheating. A question that has not been addressed is that of the thermalization of the produced particles. Interesting phenomena can also occur in this

out-of-equilibrium phase, such as the restoration of symmetries and the possibility of forming topological defects, or of undergoing a phase of inflation at low energy.

Let us mention that we have not addressed the effect of the large, time-dependent, field inhomogeneities on the space-time metric. In particular they act as a source of gravitational radiation [40]. This may be one of the only observational signatures of this phase. We refer to Ref. [41] for a general discussion.

A numerical code in free access [42], LATTICEEASY, simulates the evolution of the interacting scalar fields in an expanding space-time and can be used to study these phases of reheating.

8.7 Eternal inflation

The previous chapters have detailed the observational signatures, mostly sensitive to the dynamics of inflation during the e-folds when the observable modes become super-Hubble. However, this zone of e-folds only corresponds to a small region of the dynamics of inflation.

The discovery of the self-reproduction process, that we describe here, is a major development for inflation and cosmology. This mechanism was known for the models of old and new inflation and was then extended to models of chaotic inflation [43].

8.7.1 Heuristic argument

8.7.1.1 Mechanism

In inflationary models with large values of the inflaton (class A), quantum fluctuations can locally increase the value of the inflaton. The expansion of these regions is then quicker and their own quantum fluctuations generate new inflationary domains. This process naturally leads to a self-reproducing Universe where there are always inflating regions.

To understand this mechanism, we give a heuristic argument. We recall that regions separated by distances greater than H^{-1} can be considered to be evolving independently. So any region of size H^{-1} will be considered as an independent Universe decoupled from other regions.

Let us consider such a region of size H^{-1} for which the scalar field is homogeneous enough and has a value $\varphi \gg M_p$ and let us consider a potential of the form (8.62). In a time interval $\Delta t \sim H^{-1}$, the field classically decreases by $\Delta\varphi \sim \dot{\varphi}\Delta t \sim \dot{\varphi}/H$. The Klein–Gordon equation in the slow-roll regime then implies that $\Delta\varphi \sim -M_p^2/4\pi\varphi$. This value should be compared with the typical amplitude of quantum fluctuations, $|\delta\varphi| \sim H/2\pi \sim m\varphi/\sqrt{3\pi}M_p$.

Classical and quantum fluctuations have the same amplitude for

$$|\Delta\varphi| \sim |\delta\varphi| \iff \varphi \sim \varphi_* \equiv \frac{M_p}{2} \sqrt{\frac{M_p}{m}}. \quad (8.299)$$

We can thus distinguish between three phases in the evolution of the inflaton (Fig. 8.16): a phase during which the quantum fluctuations are of the same order (or larger) as the classical field variation, a phase in which the field is in classical slow-roll towards its minimum and a phase when the field oscillates around its minimum. Note that

$V(\varphi_*) = (m/M_p)M_p^4/8 \lesssim 10^{-6}M_p^4$ so that the first regime can be reached even at energies small compared to M_p .

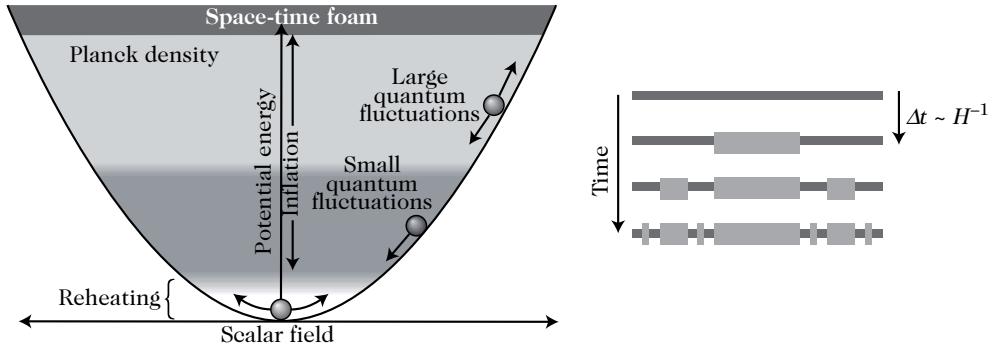


Fig. 8.16 (left): The three phases of evolution of the inflaton depend on its value compared to φ_* . For $\varphi > \varphi_*$, quantum fluctuations are of the same order as (or larger than) the classical variation of the field and the Universe is in a self-reproduction phase. For $\varphi < \varphi_*$, quantum fluctuations are negligible and the field rolls slowly towards its minimum. For $\varphi < \varphi_f$, the field oscillates around its minimum during the reheating phase. (right): During the self-reproduction phase, in a time H^{-1} , each region of size H^{-1} crumbles into roughly 20 regions that become Universes decoupled from one another. Statistically, the value of φ increases in half of them.

In the regime where $\varphi \gg \varphi_*$, $\delta\varphi \gg \Delta\varphi$. The characteristic length of the fluctuations $\delta\varphi$ generated during the time Δt is of the order of H^{-1} so that the initial volume is divided into $(\exp H\Delta t)^3 \sim 20$ independent volumes of radius H^{-1} . Statistically the value of the field in half of these regions is $\varphi + \Delta\varphi - \delta\varphi$ and $\varphi + \Delta\varphi + \delta\varphi$ in the other half. So the physical volume of the regions where the field has a value greater than φ_* is ten times larger

$$V_{t+\Delta t}(\varphi > \varphi_*) \sim \frac{1}{2} (\exp H\Delta t)^3 V_t(\varphi > \varphi_*) \sim 10 V_t(\varphi > \varphi_*). \quad (8.300)$$

The physical volume where the space is inflating therefore grows exponentially in time. The zones where the field becomes lower than φ_* enter a slow-roll phase. So they become inflationary Universes, decoupled from the rest of the Universe, with a slow-roll phase, a reheating phase and a hot Big Bang. These zones are *island Universes* (or pocket Universes) and our observable Universe is only a tiny part of such an island Universe.

8.7.1.2 Consequences for cosmological models

This scenario has important consequences for cosmology. As seen earlier, the Universe now has a very inhomogeneous structure on scales larger than H^{-1} with regions undergoing eternal inflation continuously giving rise to new zones themselves undergoing inflation. At large scales, the Universe therefore has a fractal structure with continuous

production of island Universes. Each one of these island Universes then undergoes a phase of ‘classical’ inflation, with a large number of e-folds and is thus composed of many regions of the size of our observable Universe.

The simplistic model used for this heuristic argument uses only one scalar field and a potential with a unique minimum. Realistic models in high-energy physics on the other hand involve many scalar fields. The potential of these fields can be very complex and have many flat directions and minima. So the same theory can have different vacua that correspond to different schemes of symmetry breaking. Each of these vacua can lead, at low energy, to physically different laws.

Due to exploration of this *landscape* by quantum fluctuations, the Universe finds itself divided into many regions with different physical laws (for instance, different values of the fundamental constants).

If this vision of the primordial Universe is correct, physics alone cannot provide a complete explanation for all the properties of our observable Universe since the same physical theory can generate vast regions with very different properties. So our observable Universe would have the properties it has not because the other possibilities are impossible or improbable, but simply because a Universe with such properties allows for a life similar to ours to appear.

Eternal inflation thus offers a framework to apply the anthropic principle since the self-reproduction mechanism makes it possible to generate Universes with different properties and to explore all possible vacua in the theory. Moreover, if the conditions required for the appearance of life exist in a domain of the size of our observable Universe, then they also exist in a much larger region since, in the context of eternal inflation, our Universe originates from a phase of inflation with a large number of e-folds.

This approach is used more and more to address the question of the value of the fundamental constants, which corresponds, as will be seen in Chapter 13, to a choice of vacuum in string theory. In this case, the complete space has more than 4 dimensions, and the numerous parameters necessary to describe the geometry and topology of the space of extra dimensions are scalar fields that explore what is called the *landscape* of the vacua of string theory.

We thus have a tool to address the likelihood of our Universe and of its physical properties. This framework also allows us to address questions beyond the origin of the properties of our Universe, thus defining the limitation of what we will be capable of explaining.

8.7.2 Stochastic approach

The best tool to address the description of this phase of eternal inflation is the stochastic approach [44,45]. This approach takes into account quantum fluctuations and incorporates them into the Klein–Gordon equation, hence transforming it into a Langevin equation. In some cases, this equation can be written as a Fokker–Planck equation for the distribution function of the field.

8.7.2.1 Field decomposition

In the stochastic approach, we first introduce a smoothing function $W_R(\mathbf{x})$ where R is a comoving scale and the operator φ is then decomposed into two parts, $\bar{\varphi}$ containing the long-wavelength modes and $\delta\varphi$ containing the short-wavelength modes. Since the physical smoothing scale is of the order of H^{-1} , we set

$$R = (\epsilon aH)^{-1}, \quad (8.301)$$

where ϵ is a constant numerical coefficient. If W_R is spherically symmetric and $w(kR)$ its Fourier transform, then

$$\varphi(\mathbf{x}, t) = \bar{\varphi}(\mathbf{x}, t) + \delta\varphi(\mathbf{x}, t), \quad (8.302)$$

with

$$\bar{\varphi}(\mathbf{x}, t) = \int \frac{d^3k}{(2\pi)^{3/2}} w(kR) [\varphi_{\mathbf{k}}(t) \hat{a}_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} + \varphi_{\mathbf{k}}^*(t) \hat{a}_{\mathbf{k}}^\dagger e^{-i\mathbf{k}\cdot\mathbf{x}}], \quad (8.303)$$

so that the Fourier components of the smoothed field are $\bar{\varphi}_{\mathbf{k}} = \varphi_{\mathbf{k}} w(kR)$ and $\delta\varphi_{\mathbf{k}} = [1 - w(kR)]\varphi_{\mathbf{k}}$.

8.7.2.2 The Langevin equation

In the slow-roll regime, the field $\bar{\varphi}$ satisfies the Klein–Gordon equation

$$\dot{\bar{\varphi}}(\mathbf{x}, t) = -\frac{1}{3H} V'(\bar{\varphi}) + \xi_Q(\mathbf{x}, t), \quad (8.304)$$

obtained by inserting the decomposition (8.302) into the Klein–Gordon equation. ξ_Q is a quantum noise describing the effect of the short wavelengths on the evolution of the smoothed field. The Klein–Gordon equation tells us that $\dot{\xi}_Q = -\ddot{\varphi} - V''(\bar{\varphi})\delta\varphi/3H$. If $\epsilon \ll 1$, the first term dominates and

$$\xi_Q(\mathbf{x}, t) = -\dot{\delta\varphi} = - \int \frac{d^3k}{(2\pi)^{3/2}} (kHR) w'(kR) [\varphi_{\mathbf{k}}(t) \hat{a}_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} + \text{h.c.}], \quad (8.305)$$

where we have used $\dot{\delta\varphi}_{\mathbf{k}} = -\dot{\varphi}_{\mathbf{k}}(1 - w) + \dot{w}\varphi_{\mathbf{k}} \simeq \dot{w}\varphi_{\mathbf{k}}$ when $\epsilon \ll 1$. We have also expressed \dot{w} as $\dot{w} = k\dot{R}w' = -RHw'$, where the prime represents a derivative with respect to the argument of the function. The correlation function of the noise $\langle \xi_Q(\mathbf{x}, t)\xi_Q(\mathbf{x}', t') \rangle = \langle 0|\xi_Q(\mathbf{x}, t)\xi_Q(\mathbf{x}', t')|0 \rangle$, is thus

$$\begin{aligned} \langle \xi_Q(\mathbf{x}, t)\xi_Q(\mathbf{x}', t') \rangle &= \int \frac{k^3 dk}{2\pi^2} \frac{\sin kr}{r} [HR(t)]w'[kR(t)] \\ &\quad [HR(t')]w'[kR(t')]\varphi_k(t)\varphi_k^*(t'), \end{aligned} \quad (8.306)$$

after integration over angles.

Although φ is a quantum operator, we can introduce a classical stochastic field, ϕ , the dynamics of which is described by the Langevin equation where the noise ξ_Q is

replaced by a stochastic noise ξ with a correlation function of the form (8.306). The evolution of the inflaton, smoothed out on super-Hubble scales, is then dictated by

$$\dot{\phi}(\mathbf{x}, t) = -\frac{V'(\phi)}{3H(\phi)} + \xi(\mathbf{x}, t), \quad \langle \xi(\mathbf{x}, t)\xi(\mathbf{x}', t') \rangle = \langle \xi_Q(\mathbf{x}, t)\xi_Q(\mathbf{x}', t') \rangle. \quad (8.307)$$

The general study of the Langevin equation requires us to know the function $H(\phi)$ and to determine the properties of the noise, which themselves depend on the dynamics of inflation and on the choice of the function w . For example, in de Sitter space, the noise correlation function takes the form

$$\langle \xi(\mathbf{x}, t)\xi(\mathbf{x}', t') \rangle = \frac{H^4\eta\eta'}{4\pi^2 r\epsilon^2} \int dk \sin(kr) w'\left(\frac{-k\eta}{\epsilon}\right) w'\left(\frac{-k\eta'}{\epsilon}\right) (1 + ik\eta)(1 - ik\eta') e^{-ik(\eta-\eta')}. \quad (8.308)$$

Note that when the window W_R is a Heaviside function then $w'(kR) = \delta(kR - 1)$ so that the correlation function becomes proportional to $\delta[R(t) - R(t')] \propto \delta(t - t')$. For instance, in the de Sitter case, one can check that

$$\langle \xi(\mathbf{x}, t)\xi(\mathbf{x}', t') \rangle = \frac{H^3}{4\pi^2} \frac{\sin(\epsilon a H r)}{\epsilon a H r} \delta(t - t'). \quad (8.309)$$

Restricting ourselves to this hypothesis for the window function, one can show in a general way that the Langevin equation takes the form

$$\dot{\phi}(\mathbf{x}, t) = -\frac{V'(\phi)}{3H(\phi)} + \frac{H^{3/2}(\phi)}{2\pi} \xi(\mathbf{x}, t), \quad \langle \xi(\mathbf{x}, t)\xi(\mathbf{x}', t') \rangle = \delta(t - t'). \quad (8.310)$$

This equation describes the field evolution as the combination of a random walk induced by non-correlated noise and a classical drift under the effect of the force $F(\phi) = -V'(\phi)/3H(\phi)$. So during an interval Δt , $\langle (\Delta\phi - F(\phi)\Delta t)^2 \rangle = \sigma^2(\phi)\Delta t$ with $\sigma(\phi) = H^{3/2}/2\pi$. The dependence of the result on the choice of the window function is discussed in detail in Ref. [45]. Note at this point that the transition to (8.310) is highly non-trivial since the geometry (H), now seen as a function of ϕ , is also described by random variables (in particular see Ref. [46] and references therein for discussions of this point).

8.7.2.3 Fokker–Planck equation

From a Langevin equation of the form (8.310), one can deduce a Fokker–Planck equation for the probability distribution function, $P(\phi, t) = P[\phi(\mathbf{x}, t) = \phi]$. The value of this function at the time $t + \Delta t$ is related to its value at the time t through the Green's function, $G(\phi, t|\phi', t')$ as

$$P(\phi, t + \Delta t) = \int G(\phi, t|\phi', t + \Delta t)P(\phi', t)d\phi'. \quad (8.311)$$

The computation of the Green's function in a small time interval requires us to write a discrete version of the Langevin equation (8.310). There is no unique way to perform this since the stochastic term can be evaluated at various times in the interval $[t, t+\Delta t]$,

$$\phi(t + \Delta t) = \phi(t) + \sum_j w_j \left\{ F[\phi_j(t)]\Delta t + \sigma[\phi_j(t)]\sqrt{\Delta t} \right\},$$

where $\phi_j(t) = \beta_j \phi(t) + (1 - \beta_j)\phi(t + \Delta t)$ with $\sum w_j = 1$ and $\beta = \sum w_j \beta_j$. Expanding this equation to linear order in Δt , we obtain

$$\phi(t + \Delta t) = \phi(t) + \left\{ F[\phi(t)] + \frac{1}{2}(1 - \beta) \frac{\partial \sigma^2}{\partial \phi} [\phi(t)] \right\} \Delta t + \sigma[\phi(t)]\sqrt{\Delta t}.$$

The different approximations of the Langevin equation in the interval Δt thus reduce to a modification of the force F . There are two frequent choices, $\beta = 1$ (Itô's version) and $\beta = 1/2$ (Stratanovich's version) that come down to evaluating the noise at the beginning or the middle of the time interval. Such an expansion then allows for the integration of (8.311) to second order in $\sqrt{\Delta t}$ [47].

The probability distribution function then satisfies the Fokker–Planck equation

$$\frac{\partial P}{\partial t} = \frac{\partial J(\phi)}{\partial \phi}, \quad J(\phi) = -F(\phi)P + \frac{1}{2}\sigma^{2(1-\beta)} \frac{\partial}{\partial \phi} (\sigma^{2\beta} P). \quad (8.312)$$

In our particular case, this equation takes the form

$$\frac{\partial P}{\partial t} = \frac{\partial}{\partial \phi} \left\{ \frac{V'(\phi)}{3H(\phi)} P(\phi, t) + \frac{H^{3(1-\beta)}(\phi)}{8\pi^2} \frac{\partial}{\partial \phi} [H^{3\beta}(\phi)P(\phi, t)] \right\}. \quad (8.313)$$

In the form (8.312), the Fokker–Planck equation appears as a continuity equation that follows from the conservation of probability.

The mean value of any function of the quantum operator $\bar{\varphi}$, $\langle f(\bar{\varphi}) \rangle$, can then be obtained as

$$\langle f(\bar{\varphi}) \rangle = \int f(\phi)P(\phi, t)d\phi, \quad (8.314)$$

where the mean value is now computed from the stochastic field ϕ .

8.7.2.4 Discussion

The Fokker–Planck equation allows for the study of the dynamics on scales larger than H^{-1} over which the Universe is very inhomogeneous. Note that the Fokker–Planck and Langevin equations are only equivalent when the noise is not correlated. We also stress that we have neglected metric perturbations. A more detailed version of stochastic inflation can be obtained in Ref. [48].

The stationary solution of (8.312) is given by

$$P(\phi, t) = N \exp \left[-\frac{8\pi^2 V(\phi)}{3H^4} \right] V^{-3\beta/2},$$

where N is a normalization factor. The dependence on the parameter β is hence sub-exponential. The solution therefore approaches asymptotically the distribution function

$$P_{\text{eq}}(\phi) = N \exp \left[-\frac{8\pi^2 V(\phi)}{3H^4} \right].$$

As an example, if $V = \lambda\phi^4/4$, then

$$P_{\text{eq}}(\phi) = \left(\frac{32\pi^2 \lambda}{3} \right)^{1/4} \frac{e^{-2\pi^2 \lambda \phi^4 / 3H^4}}{\Gamma(\frac{1}{4}) H}. \quad (8.315)$$

The distribution function of ϕ is not Gaussian and satisfies $\langle \phi^2 \rangle = \sqrt{3/2\pi^2} [\Gamma(3/4)/\Gamma(1/4)] (H^2/\sqrt{\lambda})$ and $\langle \phi^4 \rangle / \langle \phi^2 \rangle^2 - 3 \sim -0.812$.

It is also interesting to introduce the probability $P(\phi, t|\chi)$ for $\phi(t)$ to have the value ϕ at time t if $\phi = \chi$ initially. One can show [44] that this conditional probability distribution satisfies the adjoint equation of (8.313) called the inverse Kolmogorov equation,

$$\frac{\partial P(\phi, t|\chi)}{\partial t} = \frac{H^{3(1-\beta)}(\chi)}{8\pi^2} \frac{\partial}{\partial \chi} \left[H^{3\beta}(\chi) \frac{\partial P(\phi, t|\chi)}{\partial \chi} \right] - \frac{V'(\chi)}{3H(\chi)} \frac{\partial P(\phi, t|\chi)}{\partial \chi}. \quad (8.316)$$

The stationary solution of this equation is

$$P_{\text{eq}}(\phi, \chi) = \exp \left[\frac{3M_p^4}{8V(\phi)} \right] \exp \left[-\frac{3M_p^4}{8V(\chi)} \right]. \quad (8.317)$$

Unfortunately, if $V = 0$ at its minimum, this distribution function is not normalizable. This is due to the fact that a stationary state exists only if the stochastic diffusion, which increases ϕ , can compensate the classical drift towards the minimum of the potential. For $\phi < \phi_f$, the stochastic contribution is no longer there. So no physically acceptable solution can be stationary.

To recover stationarity, we should introduce the probability distribution function $P_{\text{phys}}(\phi, t|\chi)$, which is the analogue of $P_{\text{eq}}(\phi, t|\chi)$ defined above but in terms of physical volumes instead of comoving volumes. During an interval dt , the physical volume of a given region grows by an amount $3H(\phi)dt$ so that the evolution equation of $P_{\text{phys}}(\phi, t|\chi)$ is

$$\frac{\partial P_{\text{phys}}}{\partial t} = \frac{\partial}{\partial \phi} \left\{ \frac{V'(\phi)}{3H(\phi)} P_{\text{phys}} + \frac{H^{3(1-\beta)}(\phi)}{8\pi^2} \frac{\partial}{\partial \phi} [H^{3\beta}(\phi) P_{\text{phys}}] \right\} + 3H(\phi) P_{\text{phys}}$$

or equivalently,

$$\frac{\partial P_{\text{phys}}}{\partial t} = \frac{H^{3(1-\beta)}(\chi)}{8\pi^2} \frac{\partial}{\partial \chi} \left[H^{3\beta}(\chi) \frac{\partial P_{\text{phys}}}{\partial \chi} \right] - \frac{V'(\chi)}{3H(\chi)} \frac{\partial P_{\text{phys}}}{\partial \chi} + 3H(\chi) P_{\text{phys}}.$$

If we only consider the regions of space for which the density is lower than the Planck density, these equations then have a stationary solution of the form [44],

$P_{\text{phys}}(\phi, t | \chi) \sim \exp(cM_{\text{P}}t)A(\phi)B(\chi)$. If $V \ll M_{\text{P}}^4$, we recover the result that $B(\chi) \propto \exp[-3M_{\text{P}}^4/8V(\chi)]$ and the normalized probability to find a given volume in a state with a given value of ϕ is $\tilde{P}_{\text{phys}}(\phi, t | \chi) \sim A(\phi)B(\chi)$.

This result has an important consequence. In the heuristic argument, we had only one region of size H^{-1} with $\varphi \gg \varphi_*$. However, if we start with a large number of regions, even with $\varphi < \varphi_*$, since the probability (8.317) is non-zero, some of them can enter a phase of eternal inflation.

This also implies that any hypothesis on the initial state of the Universe becomes completely disconnected from observations. If a region of the Universe enters a self-reproducing phase, it produces an infinite number of island Universes with properties approaching that of a stationary state, independently of the initial conditions. We stress the difficulty encountered in clearly defining probabilities and the notion of measure. Nonetheless, the probability for inflation to start becomes completely irrelevant as soon as it does not completely vanish.

Moreover, it seems that the Universe is not eternal in the past and that it still has an initial singularity [49], in the sense that space-time is geodesically incomplete in the past (a space is geodesically complete as soon as all the geodesics of this space can be continued indefinitely). This conclusion relies on the hypothesis that the weak energy condition is satisfied, this condition can, however, be violated by quantum fluctuations.

8.8 Extensions

The previous framework is the minimal framework for inflation. We now present some extensions to show the richness of the inflationary paradigm. These extensions allow us to address questions on the robustness of the predictions that we have established.

8.8.1 Multifield inflation

In single-field inflationary models, the inflaton controls both the expansion of the Universe and the generation of density perturbations and gravitational waves. This explains the existence of a consistency relation.

High-energy theories (Chapter 10) contain legions of scalar fields. If some of these fields are light, they will develop super-Hubble correlations. In the framework of multifield inflation one can extend the space of possibilities for inflation, simply through straightforward ‘division of labour’ between the inflaton and auxiliary fields. In particular, additional fields make it possible to conceive the generation of isocurvature perturbations, to extend the consistency relation, to modify the spectral index and to generate non-Gaussianities.

8.8.1.1 Friedmann and Klein–Gordon equations

To start with, let us look at how the field equations are modified when there are many scalar fields. Friedmann’s equations become

$$H^2 = \frac{4\pi G_{\text{N}}}{3} \left[\sum_i^N \dot{\varphi}_i^2 + 2V(\varphi_1, \dots, \varphi_N) \right], \quad \dot{H} = -4\pi G_{\text{N}} \sum_i^N \dot{\varphi}_i^2 \quad (8.318)$$

and the Klein–Gordon equation for each field takes the form

$$\ddot{\varphi}_i + 3H\dot{\varphi}_i = -\frac{\partial V}{\partial \varphi_i}. \quad (8.319)$$

These N equations are coupled if V is not the sum of N potentials, i.e. of the form $V(\varphi_1, \dots, \varphi_n) = \sum_{i=1}^n V_i(\varphi_i)$. Note, however, that even in the absence of explicit interactions in the scalar field Lagrangian, the fields are still gravitationally coupled and, in particular, the Hubble damping that enters the individual Klein–Gordon equations involves the sum over all the fields.

Let us consider, for simplicity, the case of two fields. Both fields can be decomposed into a part tangent to the trajectory in field space, σ , and a part perpendicular, s ,

$$\begin{pmatrix} \dot{\sigma} \\ \dot{s} \end{pmatrix} = \mathcal{M}(\theta) \begin{pmatrix} \dot{\varphi}_1 \\ \dot{\varphi}_2 \end{pmatrix}, \quad \mathcal{M}(\theta) = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}, \quad (8.320)$$

where the angle θ is defined by

$$\cos \theta = \frac{\dot{\varphi}_1}{\sqrt{\dot{\varphi}_1^2 + \dot{\varphi}_2^2}}, \quad \sin \theta = \frac{\dot{\varphi}_2}{\sqrt{\dot{\varphi}_1^2 + \dot{\varphi}_2^2}}. \quad (8.321)$$

Relation (8.320) implies that s is constant and the Klein–Gordon equation reduces to

$$\ddot{\sigma} + 3H\dot{\sigma} + U_{,\sigma} = 0, \quad \begin{pmatrix} U_{,\sigma} \\ U_{,s} \end{pmatrix} = \mathcal{M}(\theta) \begin{pmatrix} V_{,1} \\ V_{,2} \end{pmatrix}. \quad (8.322)$$

The Friedmann equation takes the form

$$H^2 = \frac{4\pi G_N}{3} [\dot{\sigma}^2 + 2V(\sigma, s = \text{const.})], \quad (8.323)$$

since the value of s is a constant, by construction, along the trajectory. The fields σ and s are the adiabatic and isocurvature components of (φ_1, φ_2) .

8.8.1.2 Examples

As an example of the dynamics of multifield inflation, let us mention assisted inflation [50] in which one considers n scalar fields with steep exponential potentials

$$V = \sum_i U_i \exp \left(-\frac{\sqrt{8\pi}\lambda_i \varphi_i}{M_P} \right).$$

Each scalar field potential is too steep to drive inflation on its own if $\lambda_i^2 > 2$ but the additional damping effect due to the collective presence of the scalar field in the Friedmann equation leads to a particular power-law solution, $a \propto t^p$ with $p = 2/\lambda^2$ where the combined fields have an effective potential $V \propto \exp(-\sqrt{8\pi}\lambda\sigma/M_P)$. Thus $\lambda \rightarrow 0$ when the number of fields is increasing ($n \rightarrow \infty$) and we can have slow-roll inflation even though each $\lambda_i^2 > 2$.

Another topical example is Nflation [51] in which the effective potential is of the form

$$V = \frac{1}{2} \sum_i m_i^2 \varphi_i^2.$$

The case $n = 1$ is indeed our standard chaotic inflationary model. But to get inflation, we have seen that the initial value of the scalar field should be several Planck masses. With many fields, the collective dynamics can yield inflation even for sub-Planckian initial values of the fields.

8.8.1.3 Perturbation equations

Using the definitions (8.126) and (8.127), the perturbation equations (8.129)–(8.131) take the form

$$\dot{\Phi} + H\Phi = 4\pi G_N(\dot{\varphi}_1\chi_1 + \dot{\varphi}_2\chi_2) \quad (8.324)$$

$$3\dot{\Phi} + (3H^2 + \dot{H})\Phi - \frac{\Delta}{a^2}\Phi = -4\pi G_N(V_{,1}\chi_1 + \dot{\varphi}_1\chi_1 + V_{,2}\chi_2 + \dot{\varphi}_2\chi_2) \quad (8.325)$$

$$\ddot{\chi}_i + 3H\dot{\chi}_i - \frac{\Delta}{a^2}\chi_i + V_{,ij}\chi_j = -2V_{,i}\Phi + 4\dot{\varphi}_i\Phi, \quad (8.326)$$

where we have used $\Psi = \Phi$. By construction, δs is gauge invariant so that $\chi_s = Q_s = \delta s$, while $\dot{Q}_\sigma = \chi_\sigma + \dot{\sigma}\Phi/H$. Using

$${}^t\dot{\mathcal{M}} = -\dot{\theta} {}^t\dot{\mathcal{M}} J, \quad {}^t\ddot{\mathcal{M}} = -\ddot{\theta} {}^t\dot{\mathcal{M}} J - \dot{\theta}^2 {}^t\dot{\mathcal{M}}, \quad J \equiv \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad (8.327)$$

the Klein–Gordon equation (8.326) takes the form

$$\begin{aligned} \frac{d^2}{dt^2} \begin{pmatrix} \chi_\sigma \\ \delta s \end{pmatrix} + 3H \frac{d}{dt} \begin{pmatrix} \chi_\sigma \\ \delta s \end{pmatrix} + \left(-\frac{\Delta}{a^2} - \dot{\theta}^2 + \mathcal{M}V_{ij} {}^t\mathcal{M} \right) \begin{pmatrix} \chi_\sigma \\ \delta s \end{pmatrix} = \\ 2\dot{\theta}J \frac{d}{dt} \begin{pmatrix} \chi_\sigma \\ \delta s \end{pmatrix} + (\ddot{\theta} + 3H\dot{\theta})J \begin{pmatrix} \chi_\sigma \\ \delta s \end{pmatrix} - 2\Phi \begin{pmatrix} U_{,\sigma} \\ U_{,s} \end{pmatrix} + 4\Phi \begin{pmatrix} \dot{\sigma} \\ 0 \end{pmatrix}, \end{aligned} \quad (8.328)$$

and the Poisson equation becomes

$$-\frac{\Delta}{4\pi G_N a^2}\Phi = 2\dot{\theta}\dot{\sigma}\delta s + \ddot{\sigma}\chi_\sigma - \dot{\sigma}(\dot{\chi}_\sigma - \dot{\sigma}\Phi), \quad (8.329)$$

with $\dot{\theta} = -U_{,s}/\dot{\sigma}$. Since the potential U satisfies

$$\begin{pmatrix} U_{,\sigma\sigma} & U_{,\sigma s} \\ U_{,s\sigma} & U_{,ss} \end{pmatrix} \equiv \mathcal{M}V_{ij} {}^t\mathcal{M}, \quad (8.330)$$

we infer the coupled system for the perturbations

$$\ddot{\chi}_\sigma + 3H\dot{\chi}_\sigma + \left(-\frac{\Delta}{a^2} - \dot{\theta}^2 + U_{,\sigma\sigma} \right) \chi_\sigma = -2\Phi U_{,\sigma} + 4\dot{\sigma}\Phi + 2(\dot{\theta}\delta s) - 2\frac{\dot{\theta}}{\dot{\sigma}}U_{,\sigma}\delta s \quad (8.331)$$

$$\ddot{\delta}s + 3H\dot{\delta}s + \left(-\frac{\Delta}{a^2} - \dot{\theta}^2 + U_{,ss} \right) \delta s = -\frac{\dot{\theta}}{\dot{\sigma}}\frac{\Delta}{2\pi G_N a^2}\Phi. \quad (8.332)$$

Equation (8.331) translates into an evolution equation for Q_σ

$$\ddot{Q}_\sigma + 3H\dot{Q}_\sigma - \left[\frac{\Delta}{a^2} + \dot{\theta}^2 - U_{,\sigma\sigma} + \frac{8\pi G_N}{a^3} \left(\frac{a^3\dot{\sigma}^2}{H} \right) \right] Q_\sigma = 2\frac{d(\dot{\theta}\delta s)}{dt} - 2\left(\frac{U_{,\sigma}}{\dot{\sigma}} + \frac{\dot{H}}{H} \right) \dot{\theta}\delta s. \quad (8.333)$$

8.8.1.4 Adiabatic and isocurvature modes

Defining the curvature perturbations as in (8.151), we find that

$$\mathcal{R} = C + \frac{H}{\rho + P}(\dot{\varphi}_1\delta\varphi_1 + \dot{\varphi}_2\delta\varphi_2) = C + \frac{H}{\dot{\sigma}}\delta\sigma = \frac{H}{\dot{\sigma}}Q_\sigma. \quad (8.334)$$

This equation can also take the form

$$\mathcal{R} = \cos^2\theta\mathcal{R}_{\varphi_1} + \sin^2\theta\mathcal{R}_{\varphi_2}. \quad (8.335)$$

So σ plays the role of the inflaton and its perturbations generate curvature perturbations. The fluctuations of s do not affect \mathcal{R} ; s is an isocurvature mode.

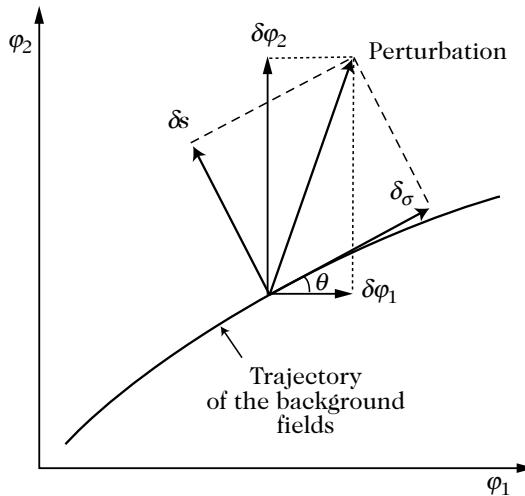


Fig. 8.17 Illustration of the scalar field decomposition into an adiabatic perturbation, tangent to the trajectory, and an isocurvature component.

\mathcal{R} is now no longer conserved (Chapter 5). Defining the entropy as

$$\mathcal{S} = H \left(\frac{\delta P}{\dot{P}} - \frac{\delta \rho}{\dot{\rho}} \right) = \frac{H}{\dot{P}} \delta P_{\text{nad}},$$

then for the super-Hubble modes, (5.150) teaches us that

$$\dot{\mathcal{R}} = -\frac{H}{\dot{H}} \frac{\Delta}{a^2} \Phi - 3Hc_s^2 \mathcal{S}. \quad (8.336)$$

We therefore find that

$$\dot{\mathcal{R}} = -\frac{H}{\dot{H}} \frac{\Delta}{a^2} \Phi + 2c_s^2 \frac{H}{\dot{\sigma}} \dot{\theta} \delta s. \quad (8.337)$$

As long as $\dot{\theta} = 0$, the two modes of perturbations are decoupled: δs behaves as a test scalar field evolving in an unperturbed space-time and the metric perturbations are only coupled to σ . If $\dot{\theta} \neq 0$, the isocurvature and adiabatic modes can mix and \mathcal{R} is no longer constant, even for super-Hubble modes.

8.8.1.5 Correlated isocurvature modes

This framework offers a rich phenomenology that we cannot give in detail.

One of the most interesting aspects relates to the isocurvature perturbations. As seen earlier, a scalar field only develops super-Hubble perturbations if it is light. Equation (8.332) tells us that the effective mass of the field s is $m_s^2 = U_{ss} + 3\dot{\theta}^2$, so that super-Hubble entropy perturbations will only be produced if

$$U_{ss} + 3\dot{\theta}^2 < \frac{3}{2}H^2. \quad (8.338)$$

Assuming that both fields are light and that initially $\dot{\theta} = 0$, which is the case during the slow-roll regime, the fields will be decoupled and each of them will generate super-Hubble fluctuations, $Q_i|_{k=aH} = (H_k \sqrt{2k^3}) e_i(\mathbf{k})$, where e_i is a random variable satisfying $\langle e_i(\mathbf{k}) e_j(\mathbf{k}') \rangle = \delta_{ij} \delta^{(3)}(\mathbf{k} - \mathbf{k}')$. Relation (8.320) implies that σ and s satisfy equivalent relations and are proportional to two random variables, e_σ and e_s , which are uncorrelated [since they are related to e_1 and e_2 by (8.320)]. During the later evolution of the super-Hubble modes, (8.331) has a source term proportional to δs if $\dot{\theta} \neq 0$. So, during this evolution, metric fluctuations develop a component proportional to e_s . We infer that the curvature and entropy perturbations can be correlated,

$$\langle \mathcal{R} \mathcal{S} \rangle \neq 0, \quad (8.339)$$

or in an equivalent way, $\langle Q_\sigma \delta s \rangle \neq 0$. To go further, we should specify how both fields decay in order to determine what kind of isocurvature perturbations are created. An interesting example [52] considers a model of double inflation [53] with potential $V = m_1^2 \varphi_1^2/2 + m_2^2 \varphi_2^2/2$ and assumes that φ_1 decays into cold dark matter and φ_2 into radiation.

In the radiation era, ‘primordial’ curvature and entropy perturbations are given by

$$\begin{pmatrix} \mathcal{R}_{RDU} \\ \mathcal{S}_{RDU} \end{pmatrix} = \begin{pmatrix} 1 & T_{\mathcal{R}\mathcal{S}} \\ 0 & T_{SS} \end{pmatrix} \begin{pmatrix} \mathcal{R}_* \\ \mathcal{S}_* \end{pmatrix}. \quad (8.340)$$

In the transfer matrix, $T_{\mathcal{R}\mathcal{R}} = 1$ since the curvature perturbation is conserved for all strictly adiabatic perturbations and $T_{\mathcal{S}\mathcal{R}} = 0$ as adiabatic perturbations cannot generate any isocurvature perturbations.

The power spectrum of the curvature perturbations is thus the sum of two components that have different spectra. Indeed, in the slow-roll regime, the spectrum of \mathcal{S} is determined by that of δs and thus depends on its mass and on the slow-roll parameters, whereas that of \mathcal{R} depends on the slow-roll parameters only.

8.8.1.6 Conclusion for the predictions of inflation

In this framework, two of the predictions from single-field inflation are modified: (1) perturbations are no longer adiabatic and (2) the spectrum of the scalar modes is no longer a nearly scale-invariant power law. We can also show that the consistency relation (8.239) is extended to a more general relation [54]. Defining

$$r_\zeta = \frac{A_{\text{T}}^2}{A_\zeta^2}, \quad r_c = \frac{A_{\zeta S}^2}{A_\zeta A_S^2}, \quad (8.341)$$

with $A_{\zeta S}^2$ defined by $A_{\zeta S}^2 = 2\langle \zeta S \rangle / 5$, the consistency relation becomes

$$r_\zeta = -\frac{n_{\text{T}}}{2}(1 - r_c^2). \quad (8.342)$$

When $r_c = 0$, the modes are completely decoupled and we recover the usual relation. Note that this implies that $r_\zeta \leq -n_{\text{T}}/2$, with equality reached only in the case of single-field inflation.

8.8.1.7 Curvaton scenario

This scenario [55–57] relates the origin of the primordial fluctuations not to the inflaton but to the quantum fluctuations of an auxiliary field, χ , now called the *curvaton* (Ref. [55] did not name the extra field).

The field χ should first be light during inflation ($m_\chi < H_*$) in order to develop super-Hubble fluctuations. Its power spectrum then has a spectral index given by (8.120). Next, the curvature perturbations generated by the inflaton should be negligible, which is possible if the energy scale of inflation is low enough, $H_* \varepsilon_*^{-1/2} < 10^{-5} M_{\text{P}}$.

As soon as the Hubble constant becomes smaller than the mass of the curvaton, the latter starts oscillating and behaves as a pressureless fluid. This can happen after the end of inflation. If $H_* \ll \chi_*$, the density contrast of this fluid is $\delta\rho_\chi/\rho_\chi = 2\delta\chi/\chi$, so that $\mathcal{P}_\delta = 2\mathcal{P}_\chi/\chi^2 = H_*/\pi\chi_*$. In this phase, the Universe is composed of a mixture of matter and radiation. The quantities ζ_{BST} for radiation and χ are separately conserved since the two fluids do not interact. The total curvature perturbation is thus

$$\zeta_{\text{BST}} = \frac{4\rho_r \zeta_r + 3\rho_\chi \zeta_\chi}{4\rho_r + 3\rho_\chi}.$$

Before the beginning of the oscillatory phase, $\zeta_{\text{BST}} = \zeta_r$, which was assumed to be negligible. So,

$$\zeta_{\text{BST}} = \frac{r}{4+3r} \delta_\chi, \quad r = \frac{\rho_r}{\rho_\chi}. \quad (8.343)$$

Depending on whether χ decays before or after having dominated the matter content of the Universe, $\zeta = \delta_\chi/3$ or $\zeta = r\delta_\chi/4$.

Finally, the field χ should decay and transfer its perturbations to radiation or matter. This imposes a constraint on the mass of the curvaton since this decay must have occurred before primordial nucleosynthesis. In the simplest version of this scenario, all the perturbations of the curvaton are transformed into adiabatic perturbations.

Unlike inflation, primordial density fluctuations are sensitive to the physics after they have become super-Hubble. Furthermore, the generation of gravitational waves is disconnected from that of the density perturbations. The coherence relation is therefore lost and $R \ll \varepsilon$.

8.8.1.8 Modulated fluctuations

Light fields can modify the predictions of inflation in another way. If the coupling term of the inflaton to the matter fields depends on the value of such a light field according to

$$\alpha \left(\frac{\chi}{M} \right) \varphi \psi \bar{\psi},$$

then the decay rate of the inflaton during the reheating phase can vary on super-Hubble scales [58, 59]. Since $\Gamma_\varphi \propto \alpha^2 m_\varphi$ and the reheating temperature is $T_{\text{reh}} \sim \sqrt{\Gamma_\varphi M_p}$, we infer that this induces density fluctuations for the radiation fluid of the order of

$$\frac{\delta \rho_r}{\rho_r} \sim \frac{\delta T_{\text{reh}}}{T_{\text{reh}}} \sim \frac{\delta \alpha}{\alpha} \sim \left(\frac{\alpha'}{\alpha} \right)_{\text{reh}} \frac{\delta \chi}{M}.$$

If the metric perturbations produced during inflation are negligible, the gravitational potential during the matter era is then of the order of $\Phi = \delta \Gamma_\varphi / \Gamma_\varphi$ [59].

This scenario can also be implemented in the context of hybrid inflation [60] where the reheating phase is tachyonic. The transition between the inflationary phase and the radiation era can be considered as instantaneous. If the effective potential is of the form

$$V = \frac{1}{4} \lambda \left(\frac{\chi}{M} \right) (\sigma^2 - v^2)^2 + \frac{g^2(\chi)}{2} \varphi^2 \sigma^2 + V(\varphi),$$

the end of inflation occurs when $\varphi = \varphi_c(\chi)$. For $\varphi > \varphi_c$, $\sigma = 0$ and the system decouples into a system that describes φ coupled to metric perturbations and χ as a test field with potential $\lambda(\chi/M)v^4/4$. Using the junction conditions (8.167) and (8.168) on the surface $q = \varphi - \varphi_c(\chi)$, one can show

$$\mathcal{R}_{RDU} = \mathcal{R}_{\text{inf}} - \sqrt{\frac{4\pi}{\varepsilon}} \gamma \frac{\delta \chi}{M_p},$$

with $\gamma \equiv d\varphi_c/d\chi$ at the time of the transition. If χ is light during inflation, it then generates super-Hubble fluctuations that modulate the end of inflation. Since $m_\chi^2 =$

$(v^4/4)\partial\lambda/\partial\chi \sim H^2 M_p^2/M^2$, χ can only be light if λ does not depend on the field, the dependence of φ_c must then be inherited from g .

The gravitational wave power spectrum is not sensitive to the precise form of the transition between inflation and the radiation era. We infer that the dispersion relation takes the form

$$R = \frac{\varepsilon}{1 + 4\pi\gamma^2}. \quad (8.344)$$

Again we find that $R < \varepsilon$.

8.8.2 Non-Gaussianity

Since the primordial curvature perturbations are related to the quantum fluctuations of a scalar field that is initially in its vacuum state, they have Gaussian statistics. As long as the evolution of the perturbations is linear, no non-Gaussianity can be generated.

However, since general relativity is non-linear, we expect some deviations from Gaussianity. In particular, such deviations can be large in multifield inflation.

This question and the observational constraints are given in detail in the review [61].

8.8.2.1 Single-field inflation

Neglecting contributions from the metric perturbations, the Klein–Gordon equation for a scalar field generally takes the form

$$\ddot{\delta\varphi} + 3H\dot{\delta\varphi} + \frac{k^2}{a^2}\delta\varphi = -V''\delta\varphi - 3V''(\delta\varphi)^2 + \dots$$

In order for the field to generate non-Gaussianity during the inflationary phase, non-linear terms should be non-negligible. If the quadratic term dominates the expansion of the r.h.s. then $\delta\varphi > V''/V'''$. As seen earlier, $\delta\varphi \sim H$. While the wavelength band $[\lambda, \lambda + \Delta\lambda]$ associated with observable scales has become super-Hubble, the slow-roll parameter η_V has varied by $\Delta\eta_V \sim (8\pi G\xi_V^2 - \eta_V)\sqrt{\varepsilon_V}\Delta\varphi/M_p$. We infer that during a number of e-folds N , η_V varies by $10^5\eta_V N$, using for these modes $H/\sqrt{\varepsilon} \sim 10^{-5}M_p$.

So if the quadratic term dominates, the slow-roll regime ends rapidly. This is because the slow-roll regime imposes the potential to be very flat whereas it must have a significant curvature for the non-linear terms to dominate. This explains why initial perturbations remain Gaussian during the inflationary phase.

To second order in the perturbations, non-linearities can induce some non-Gaussianity, but the latter must be weak. A way to parameterize this effect is to introduce a deviation (satisfying a χ^2 statistics) that characterizes the three-point function. In Fourier space,

$$\Phi_{\mathbf{k}} = \Phi_{\text{lin}}(\mathbf{k}) + \int \frac{d^3\mathbf{k}_1 d^3\mathbf{k}_2}{(2\pi)^{3/2}} f_{\text{NL}}^{(\Phi)}(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}) \Phi_{\text{lin}}(\mathbf{k}_1) \Phi_{\text{lin}}(\mathbf{k}_2) \delta^{(3)}(\mathbf{k} - \mathbf{k}_1 - \mathbf{k}_2), \quad (8.345)$$

where $f_{\text{NL}}^{(\Phi)}(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k})$ is a function that can be computed in some models [61] and Φ_{lin} is the Bardeen potential in the linear regime. With this definition,

$$\langle \Phi_{\mathbf{k}_1}(\eta) \Phi_{\mathbf{k}_2}(\eta) \Phi_{\mathbf{k}_3}(\eta) \rangle = \frac{\delta^{(3)}(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3)}{(2\pi)^{3/2}} \left[f_{\text{NL}}^{(\Phi)}(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) P_{\Phi_{\text{lin}}}(\mathbf{k}_1) P_{\Phi_{\text{lin}}}(\mathbf{k}_2) \right. \\ \left. + (1 \rightarrow 2 \rightarrow 3) + (1 \rightarrow 3 \rightarrow 2) \right].$$

Several f_{NL} are introduced in the literature, depending on the variable they parameterize the non-Gaussianity, e.g. Φ or \mathcal{R} or $\tilde{\mathcal{R}}$ defined by $\exp 2\tilde{\mathcal{R}} = 1 + 2\mathcal{R}$. These f_{NL} are related by

$$f_{\text{NL}}^{(\Phi)} = \frac{5}{3} f_{\text{NL}}^{(\mathcal{R})}, \quad f_{\text{NL}}^{(\tilde{\mathcal{R}})} = f_{\text{NL}}^{(\mathcal{R})} + 1.$$

In a model of single-field inflation, the non-Gaussianity is induced by the second-order dynamics and this coefficient is typically of the order of [62]

$$f_{\text{NL}}^{(\tilde{\mathcal{R}})}(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) \sim \frac{1}{4}(n_s - 1) + \varepsilon g(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3), \quad (8.346)$$

where $g(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3)$ is a function such that $g \rightarrow 0$ in the limit $k_1 \ll k, k_2$ or $k_2 \ll k, k_1$. As expected f_{NL} is of the order of the slow-roll parameters.

Analysis from the WMAP data [29] can be used to establish the constraint

$$-9 < f_{\text{NL}}^{(\Phi)} < 111 \quad (8.347)$$

at 95% confidence level if f_{NL} is constant. Future experiments, such as Planck, hope to be able to reach a value of $f_{\text{NL}} \sim 5$.

In the context of single-field models, an alternative is to consider potentials that have characteristic scales for which the derivative of the potential varies very rapidly [63]. These models induce a non-Gaussianity localized at the given scale.

For more details concerning the theory of cosmological perturbations at second order and its implication on non-Gaussianity, we refer the reader to Refs. [62, 64–66].

8.8.2.2 Extensions that allow for the generation of non-Gaussianity

Various extensions of the single-field model allow for the production of important non-Gaussianity. Let us cite different sources that have been considered and we refer to Ref. [61] for a review on the state-of-the-art.

- *topological defects*: these models introduce scalar fields with a negligible contribution to the background energy. They are therefore pure perturbations. Since the energy is quadratic with respect to the field, it induces some non-Gaussianity in the perturbations. We can mention the models of χ^2 -‘seeds’ [67], of axions [68] and of topological defects [69].
- *multifield inflation*: In this case, non-linear effects can be much more important in the isocurvature direction without affecting the dynamics of inflation. The auxiliary field becomes non-Gaussian and this non-Gaussianity is then transferred to the adiabatic mode in a later phase for which $\dot{\theta} \neq 0$ [61] or contributes to an isocurvature mode [55]. Interestingly, the one-point probability distribution function of the curvature perturbation at the end of inflation can be computed exactly [70].

- *curvaton*: in the limit where $H_* \gg \chi_*$, the perturbation is larger than the unperturbed value so that the density contrast during the phase where the curvaton oscillates is $\delta = (\delta\chi^2)/\langle(\delta\chi^2)\rangle$. This regime is strongly constrained by observations, but in the intermediate case where $H_* \sim \chi_*$ the curvature perturbation (8.343) becomes [57]

$$\xi_{\text{BST}} = \frac{r}{4} \left[2 \frac{\delta\chi}{\chi} + \left(\frac{\delta\chi}{\chi} \right)^2 \right],$$

leading to $f_{\text{NL}}^{(\Phi)} = -7/6 - 5r/6 + 5/4r$ to which one should add the effects of perturbations at second order.

8.8.3 Trans-Planckian problem

The predictions of inflation are based on the use of quantum field theory in curved space-time. Even though the metric is treated in a quantum way, we have shown that at linear order, technically everything reduces to the quantization of a free scalar field in a Friedmann–Lemaître space-time.

8.8.3.1 Origin of the problem

If we consider, for concreteness, an inflationary model with a potential of the form (8.68) with $n = 4$, (8.70) implies that the value of the inflaton at the moment at which observable scales become super-Hubble is $(\varphi_*/M_p)^2 = (N_* + 1)/\pi$. We infer that the Hubble parameter is then $(H^*/M_p)^2 = \lambda_4(\varphi_*/M_p)^4 = \lambda_4(N_* + 1)^2/\pi^2$. According to (8.69), the slow-roll parameter is $\varepsilon_* = 1/(N_* + 1)$, from which we infer that

$$\frac{H_*}{\varepsilon_* M_p} = \frac{\lambda_4}{\pi^2} (N_* + 1)^3.$$

Observational constraints impose this quantity to be of the order of 10^{-10} , which fixes $\lambda_4 \sim 10^{-13}$, using $N_* \sim 60$. However, the initial condition of chaotic inflation $V(\varphi_i) \sim M_p^4$ imposes the total number of e-folds to be of the order of $N \sim 5 \times 10^8$. We infer that the wavelength of the mode corresponding to the Hubble radius today, $\ell_{H_0} \sim 10^{61} \ell_p$, had a wavelength of the order of $\ell_H \sim 10^{61} \exp(-N) \ell_p \ll \ell_p$. So the modes observable today had a sub-Planckian length at the beginning of inflation.

We can estimate the necessary number of e-folds in order for the observable scales to be sub-Planckian during inflation from (8.61). Assuming that $T_{\text{reh}} \sim V^{1/4} \sim 10^{16} \text{ GeV}$, this relation tells us that the mode corresponding to the Hubble scale today had a Planck length at around 70 e-folds before the end of inflation.

Generically, around ten e-folds beyond the required 60 e-folds are sufficient for the currently observable modes to have been sub-Planckian. It is legitimate to ask ourselves to what extent the results obtained so far are robust compared to the unknown physics at the Planck scale. As illustrated in Fig. 8.18, at the moment at which these modes become super-Planckian, the Hubble constant is well below the Planck scale, which implies that the description of space-time with a Friedmann–Lemaître space-time remains justified. The trans-Planckian problem [71] therefore only involves fluctuations and is different from the issue of the space-time description when $V(\varphi) \sim M_p^4$.

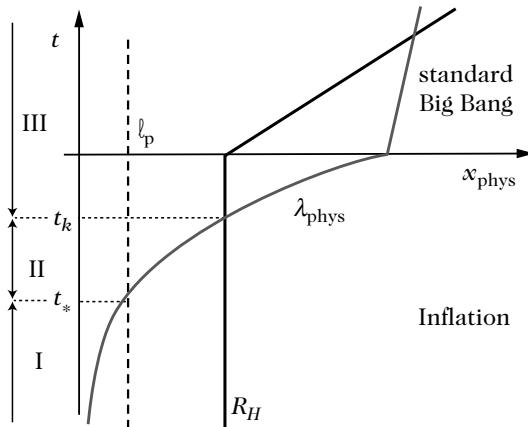


Fig. 8.18 Evolution of the physical wavelength of a given mode and of the Hubble radius as a function of time. The mode becomes super-Hubble at t_k and is sub-Planckian for $t < t_*$. In region II, the sub-Hubble mode is cis-Planckian; it is in this regime that the quantum field theory approach is under control.

Two related questions then arise. Are predictions from inflation robust with respect to this unknown physics and can this physics leave an observational signature? The main problem to answer these questions lies in the fact that this physics is so far unknown. It is therefore difficult to predict its influence on inflation. To get around this problem, the question is so far addressed by introducing *ad-hoc* modifications of quantum field theory that are supposed to describe modifications from quantum gravity. The conclusion from these studies is twofold. If the predictions from inflation are robust with respect to these modifications, then it is reasonable to think that they will also be so with respect to that induced by quantum gravity. If this is not the case, then the real theory is necessary in order to determine the exact amplitudes of the expected effects.

Among the many proposals in the literature, we describe two of them, the modification of the dispersion relation and the so-called ‘minimal’ approach.

8.8.3.2 Model with a modified dispersion relation

The initial approach to this problem follows the one developed to study the Hawking radiation of black holes. The origin of the term in k^2 in the evolution equation of v_k and μ_k can be traced to the assumption that the relation $\omega_{\text{phys}} = k_{\text{phys}}$ was satisfied. Just as the dispersion relation changes in condensed-matter physics as the wavelength becomes of the order of the inter atomic distance, we can conjecture that the dispersion relation of quantum fluctuations will be modified for physical wavelengths of the order of the Planck length. We can therefore hope to model the ‘discrete’ structure of space-time by modifying the dispersion relation into

$$k_{\text{phys}} = \omega_{\text{phys}}(k_{\text{phys}}), \quad (8.348)$$

which, in terms of comoving quantities, amounts to the modification

$$k^2 \rightarrow k_{\text{eff}}^2(k, \eta) = a^2 F^2 \left(\frac{k}{a} \right) \equiv \omega^2(k, \eta). \quad (8.349)$$

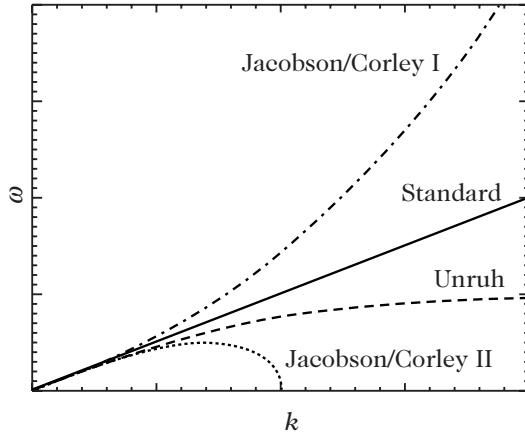


Fig. 8.19 Some of the dispersion relations considered in the literature.

The function F is linear as long as $k/a \ll k_C$, where k_C is a new scale. Figure 8.19 illustrates some dispersion relations that are used. The dispersion relation hence becomes time dependent and the evolution equation of v_k or μ_k becomes

$$\begin{aligned} v_k'' + \omega_T^2(k, \eta)v_k &= 0 \\ \omega_T^2(k, \eta) &= \left[\omega^2(k, \eta) - \frac{z''}{z} \right]. \end{aligned} \quad (8.350)$$

One can obtain a rigorous derivation of this equation from a variational principle [72]. Note that the evolution equation remains linear, Gaussianity will be preserved.

To follow the standard procedure, we ask that v_k satisfy the Wronskian normalization (8.182). We should then impose a positivity condition for the modes frequency, as these modes will then be used to define the operators \hat{a}_k^\dagger from which the Fock space can be constructed. Physically, we should impose that at the initial time η_i , the modes $k/a \gg H$ are in their fundamental state. In the usual approach, there is no ambiguity in defining the vacuum since k^2 is constant. With a modified dispersion relation, ω depends on η even at the initial time. We therefore lose the conventional definition of the vacuum.

A solution is to use the adiabatic approach [24] and to assume that the initial configuration minimizes its energy density. This energy density takes the form [72]

$$\rho = \frac{1}{4\pi a^2} \int dk k^2 \left[\left| \left(\frac{v_k}{a} \right)' \right|^2 + \omega^2 \left| \frac{v_k}{a} \right|^2 \right]. \quad (8.351)$$

Decomposing v'_k as $v'_k = x + iy$ and imposing the normalization (8.182), the energy minimization implies that

$$v_k(\eta_i) = \frac{1}{\sqrt{2\omega(k, \eta_i)}}, \quad v'_k(\eta_i) = -i\sqrt{\frac{\omega(k, \eta_i)}{2}}. \quad (8.352)$$

This prescription coincides with that obtained by choosing a locally Minkowski state when $\omega = k$. This adiabatic vacuum is associated to a solution, of so-called WKB-type [73], with positive frequency

$$v_k = \frac{1}{\sqrt{2\omega(k, \eta)}} \exp \left[-i \int_{\eta_i}^{\eta} \omega(k, u) du \right]. \quad (8.353)$$

This solution is valid as long as the adiabaticity condition

$$\left| \frac{\omega'_T}{\omega_T^2} \right| = \left| \frac{H}{\omega_{\text{phys}}} \left(1 - \frac{d \ln \omega_{\text{phys}}}{d \ln k_{\text{phys}}} \right) \right| \ll 1 \quad (8.354)$$

is satisfied.

If this adiabaticity condition is not violated between η_i and a time η_f for which the dispersion relation is linear and the mode is sub-Hubble, then the solution for $\eta > \eta_f$ is

$$v_k \simeq \frac{1}{\sqrt{2k}} \exp \left[-ik\eta - i \int_{\eta_i}^{\eta_f} \omega(k, u) du \right]. \quad (8.355)$$

The phase has no effect on the absolute value of v_k and on the observables.

So, in order for a modification to be present, the dispersion relation must be such that there is a violation of the WKB regime. In this case, the solution at η_f is of the form

$$v_k = \frac{1}{\sqrt{2\omega(k, \eta)}} \left\{ \alpha_k \exp \left[-i \int_{\eta_f}^{\eta} \omega(k, u) du \right] + \beta_k \exp \left[+i \int_{\eta_f}^{\eta} \omega(k, u) du \right] \right\}, \quad (8.356)$$

with $|\alpha_k|^2 - |\beta_k|^2 = 1$. This implies that primordial fluctuations are modulated and that everything happens as if we defined the vacuum at η_f from a non-vacuum state. In region II (Fig. 8.19) the term z''/z is negligible and $F = k/a$. We infer that the primordial spectrum becomes

$$\mathcal{P} = \mathcal{P}_0 \left\{ 1 + 2\Re \left[\beta_k (e^{i\phi_k}) \right] + \mathcal{O}(\beta^2) \right\}, \quad (8.357)$$

where \mathcal{P}_0 is the spectrum obtained using the condition (8.184) and ϕ_k is a phase. This relation is valid if

$$\sigma \equiv \frac{k_C}{H} \gg 1. \quad (8.358)$$

Expanding the solution to second order in the WKB approximation, one can show that if (8.354) is satisfied between η_i and η_f then $\beta_k \sim \exp(-\sigma)$. To conclude, if condition (8.354) is satisfied and $\sigma \gg 1$ then the predictions of inflation are not modified [74]. On the other hand, if these conditions are not satisfied, some spectrum modulations appear due to the existence of negative-frequency modes at η_f . However, this situation is constrained since a non-vacuum state is at the origin of an important gravitational backreaction that could jeopardize the entire inflationary regime.

8.8.3.3 Commutation relation and minimal approach

A second approach is based on a modification of the commutation relations of quantum mechanics [75]. Assuming that the standard commutation relations become

$$[x, p] = i\hbar(1 + \beta p^2 + \dots),$$

then the uncertainty relation transforms into

$$\Delta x \cdot \Delta p \geq \frac{\hbar}{2}(1 + \beta^2 \Delta p^2 + \dots).$$

This implies the existence of a minimal distance $\ell_{\min} = \sqrt{\beta}\hbar$. This modification has two effects. The equation of evolution of the perturbations is modified and the uncertainty relation saturates when the physical wavelength is of the order of ℓ_{\min} . We can then assume that each mode has its own creation time defined by

$$\frac{2\pi}{k}a(\eta_k) = \ell_{\min}, \quad (8.359)$$

since there is no sense in talking about fluctuations before η_k . So η_k depends on k , which should be compared with the standard case where $\eta_k \rightarrow -\infty$ for all the modes.

A minimal approach [76] attempts to evade the trans-Planckian problem completely. It assumes that the mode is created at η_k and then evolves with the standard equation (8.179). The most general state at η_k is of the form

$$v_k = \frac{c_k + d_k}{\sqrt{2\omega(k, \eta_k)}}, \quad v'_k = i(c_k - d_k)\sqrt{\frac{\omega(k, \eta_k)}{2}}, \quad (8.360)$$

with $|c_k|^2 - |d_k|^2 = 1$. The coefficients can be expanded as a function of $\sigma = H\ell_{\min}$ as

$$c_k = 1 + y\sigma + \dots \quad d_k = x\sigma + \dots$$

When $\sigma \rightarrow 0$, we recover the prescription of the usual vacuum ($c_k = 1$ and $d_k = 0$). The parameters x and y are completely free and σ is a small parameter. We can show that this induces the presence of amplitude oscillations $|x|\sigma$ in the power spectrum [77],

$$\begin{aligned} \mathcal{P}_\zeta &= \mathcal{P}_\zeta(\sigma = 0) \left\{ 1 - 2|x|\sigma \cos \left[\frac{2}{\sigma} (1 + \varepsilon + \varepsilon \ln k\ell_{\min}) + \psi \right] \right\} \\ &\quad - 2 \frac{H^2}{\pi\varepsilon M_P^2} |x|\sigma (2\varepsilon - \delta) \sin \left[\frac{2}{\sigma} (1 + \varepsilon + \varepsilon \ln k\ell_{\min}) + \psi \right] \end{aligned} \quad (8.361)$$

where ψ is a phase to compare with (8.227). The wavelength of the oscillations is $\Delta \ln k = \sigma\pi/\varepsilon$. Moreover, to avoid gravitational backreaction we should have $|x| \leq 10^4\sigma/\sqrt{\varepsilon}$.

8.8.3.4 Backreaction

As illustrated in the example of the modified dispersion relation, these models generically lead to states with $\beta_k \neq 0$ at the time η_f . So, these models roughly reduce to considering a mixture of positive and negative frequencies at this initial time.

If the observable modes are initially in a non-vacuum state, then their energy density rapidly dominates that of the Friedmann–Lemaître space. Indeed, such a mode becomes super-Hubble at $N(k) \sim 56 - \ln(hk/a_0 H_0)$ e-folds before the end of inflation. Its physical momentum is then $k_{\text{phys}} = H_{\text{inf}}$. So this mode had a momentum $k_{\text{phys}} = \exp(N_i)H_{\text{inf}}$ at the beginning of inflation, $N_i = N_{\text{tot}} - N(k)$. As seen earlier, in general $N_i \gg 1$ if the chaotic initial conditions are imposed. The energy density of this mode is of the order of $\delta\rho_k \sim n_k \exp(4N_i)H_{\text{inf}}^4/2\pi^2$ assuming that the zero-point energy has been subtracted. We infer that

$$\frac{\delta\rho_k}{\rho_{\text{crit}}} \sim n_k \exp(4N_i) \left(\frac{H_{\text{inf}}}{M_P} \right)^2. \quad (8.362)$$

To avoid any backreaction problem, this number should be smaller than unity. Fixing the inflation scale to $10^{-6}M_P$, this imposes that

$$n_k \lesssim \exp(28 - 4N_i) \iff N_i \lesssim 7. \quad (8.363)$$

This restriction is unnatural so that we should conclude that we need $n_k \ll 1$. Gravitational backreaction is therefore only avoided if these modes are initially in the vacuum or in an extremely close state. The general problem of backreaction is discussed in Ref. [78].

In the case of the modified dispersion relations, one can generically conclude that if the adiabaticity condition (8.354) has been violated during the evolution of a given mode, then when this condition is restored, this mode behaves as a mixture of positive and negative frequencies, which means that a number of quanta have been created and that their energy density is large today compared to the critical density of the Universe [72, 79].

In the trans-Planckian case, one can show that if the adiabaticity condition is violated during a period $\Delta \ln \eta = \epsilon$ then $\beta_k \propto \epsilon$ and $\alpha_k \sim 1 - \epsilon$. The energy density of the fluctuations is then $\delta\rho \sim \mathcal{O}(\epsilon^2 k_C^2)$ so that in order to avoid any backreaction problem, we should have $\epsilon k_C^2 \lesssim M_P H_{\text{inf}}$ [72].

The backreaction problem is a generic problem of trans-Planckian modifications. In general, imposing that this problem is resolved, reduces to zero the possibility of observing any specific effects.

8.9 Status of the paradigm

8.9.0.5 Generic predictions

Single-field inflationary models have robust predictions that are independent of their specific implementation.

- The observable Universe must be homogeneous and isotropic. Inflation erases any classical inhomogeneities. The observable Universe can thus be described by a Friedmann–Lemaître Universe.
- The Universe must be spatially flat. During inflation, the curvature of the Universe is exponentially suppressed. Inflation therefore predicts that $\Omega = 1$ up to the amplitude of the super-Hubble density perturbations, i.e. up to about 10^{-5} .

- Scalar perturbations are generated. Inflation finds the origin of the density perturbations in the quantum fluctuations of the inflaton that are amplified and redshifted to macroscopic scales. All the modes corresponding to observable scales today are super-Hubble at the end of inflation. Inflation predicts that these perturbations are adiabatic, have Gaussian statistics and have an almost scale invariant power spectrum. The spectral index can vary slightly with wavelength. These perturbations are coherent, which is translated into a structure of acoustic peaks in the angular power spectrum of the CMB's temperature anisotropies (Chapter 6).
- There are no vector perturbations. Vector modes are exponentially suppressed.
- Gravitational waves are generated. In the same way as for scalar modes, the gravitational waves have a quantum origin and are produced through parametric amplification. They also have Gaussian statistics and an almost scale invariant power spectrum. There is a consistency relation between their spectral index and the relative amplitude of the scalar and tensor modes.
- Any light test field generates super-Hubble fluctuations. The quantum fluctuations of any light field ($M < H$) are amplified. This field develops super-Hubble fluctuations of amplitude $H/2\pi$.

More generally, inflationary models radically change our vision of cosmology in two respects.

- Particles are produced by the preheating mechanism. The inflaton decay allows us to explain the production of particles at the end of inflation.
- Inflation is eternal. Placing inflationary models in the context of chaotic initial conditions, we obtain a very different picture of the Universe. The latter should be in eternal inflation and gives rise to island Universes.

We can conclude that inflation offers a satisfying framework to think about the Universe, and in particular about the initial conditions. It solves the cosmological problems and describes the origin of the large-scale structure.

8.9.0.6 Current constraints

Cosmological observations allow us to test the dynamics of inflation during the $8 - 10$ e-folds during which the observable scales become super-Hubble. The study of this phase is mainly based on the slow-roll approximation.

Current observations, and in particular those by WMAP [28] can be used to constrain some of the parameters of primordial the spectra. Observations from the cosmic microwave background are compatible with Gaussian, adiabatic and almost scale invariant primordial fluctuations.

Gravitational waves have not yet been detected and their contribution is constrained to be lower than 30% of the total signal. As stressed, their detection is essential to check the consistency relation, which would be a proof of inflation. The measurement of α_s may be constrained by using observation at smaller scales, such as weak lensing [80] or the Lyman- α forest.

Inflation predicts a flat Universe ($|\Omega_0 - 1| < 10^{-5}$) and current observations are compatible with $\Omega_0 = 1$. The WMAP-1 data slightly favoured a spherical universe [28] since $\Omega_0 = 1.02 \pm 0.02$. Although more recent data, leading to $\Omega_0 = 1.0049 \pm 0.013$ [29],

no longer support a closed Universe, we stress that the detection of a positive curvature, even a small one, would be in itself a major problem for inflation [81]. For this let us model inflation by a de Sitter phase with $K = +1$. The scale factor evolves as $a = a_i \cosh(\sqrt{\Lambda/3}t)$ and $\delta = \Omega - 1$ evolves as $\delta = 1/\sinh^2(\sqrt{\Lambda/3}t)$, where $t = 0$ corresponds to the minimum of the scale factor during inflation. The time at which the mode $k_0 = a_0 H_0$ becomes super-Hubble is thus given by $\sqrt{\Lambda/3}t_* = \text{arcsinh}(1/\sqrt{\delta_0})$. We infer that the maximal number of e-folds before this moment, i.e. between $t = 0$ and t_* , is

$$\Delta N_{\max} = \frac{1}{2} \ln \left(1 + \frac{1}{\delta_0} \right) \sim 2, \quad (8.364)$$

for $\delta_0 \sim 0.02$. The picture of the primordial phase would thus be strongly modified if a non-vanishing curvature was shown to be present, albeit the fact that today we have the best estimate $\Omega_0 = 1.0049 \pm 0.013$ [29]. Models of open [82] and closed [83] inflation have been proposed. We stress that the existence of a curvature induces a characteristic scale that can be associated with a deviation from a strict power law for the primordial power spectrum. The robustness with respect to a primordial shear was investigated in Ref. [84].

Finally, the causal structure of inflation has been studied and one can show that the initial singularity prevails [49]. In this sense, the theory is not complete and the problem of initial conditions is not defined.

8.9.0.7 Conclusions

Single-field inflation has robust, testable and falsifiable predictions. It is therefore a completely satisfying physical model that we should accommodate in the framework of the standard cosmological model. This model therefore includes the hot Big-Bang model, which describes the post-inflationary evolution of our Universe, and the model of inflation that describes its primordial phase.

This model is predictive and its predictions, based on general relativity and quantum field theory, are not disputed. From a technical point of view the linear analysis is thus on a solid basis. Notice, however, that these predictions are based on the slow-roll hypothesis at second order. These predictions are only relevant for around ten e-folds, i.e. a small interval of variation of the inflaton.

In this framework, we should therefore add 5 additional cosmological parameters: the amplitude, the spectral indices of the density and gravitational waves power spectra and their derivatives, that is 6 parameters minus 1 for the consistency relation. The typical amplitudes of density fluctuations and of gravitational waves are $H/\sqrt{\varepsilon}$ and H . Observations imply that the energy scale of inflation is of the order of 10^{16} GeV.

Explicit models that fit naturally in a framework of high-energy physics are yet to be constructed. The Grand Unified and supersymmetric theories have many scalar-field candidates. But we should stress that today no scalar field has yet been directly detected. Note also that this scalar field is not necessarily a fundamental field since $f(R)$ theories are dynamically equivalent to general relativity with a non-minimally coupled scalar field. So, a central question of these models is *what is the nature of the inflaton?*

540 *Inflation*

This model also has many extensions (multifield inflation, . . .) that allow us to test the robustness of the predictions. It offers, with eternal inflation, a framework to use the anthropic principle.

Other questions remain open and require deeper studies. This is the case for the trans-Planckian modes, for the perturbations backreaction and for the past completeness of inflation.

References

- [1] E. GLINER, ‘Algebraic properties of the energy-momentum tensor and vacuum-like states of matter’ *Sov. Phys. JETP*, **22**, 378, 1966.
- [2] R. BROUT, F. ENGLERT and E. GUNZIG, ‘The creation of the Universe as a Quantum phenomenon’, *Ann. Phys.* **115**, 78, 1978.
- [3] A. STAROBINSKY, ‘Spectrum of relic gravitational radiation and the early state of the Universe’, *JETP Lett.* **30**, 682, 1979.
- [4] A. GUTH, ‘Inflationary Universe: a possible solution to the horizon and flatness problems’, *Phys. Rev. D* **23**, 347, 1981.
- [5] A. D. LINDE, ‘A new inflationary scenario: a possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems’, *Phys. Lett. B* **108**, 389, 1982;
- [6] V. MUKHANOV and G. CHIBISOV, ‘Quantum fluctuations in a nonsingular Universe’, *JETP Lett.* **33**, 532, 1981.
- [7] S. HAWKING, ‘The development of irregularities in a single bubble inflationary Universe’, *Phys. Lett. B* **115**, 295, 1982; A. STAROBINSKY, ‘Dynamics of phase transition in the new inflationary scenario and generation of perturbations’, *Phys. Lett. B* **127**, 175, 1982; A. GUTH and S.-Y. PI, ‘Fluctuations in the new inflationary Universe’, *Phys. Rev. Lett.* **49**, 1110, 1982; J. BARDEEN, P. STEINHARDT and M. TURNER, ‘Spontaneous creation of almost scale-free density perturbations in an inflationary Universe’, *Phys. Rev. D* **28**, 679, 1983.
- [8] A. LINDE, ‘Chaotic inflation’, *Phys. Lett. B* **129**, 177, 1983.
- [9] A. LINDE, *Particle physics and inflationary cosmology*, Harwood Academic Publishers, 1990.
- [10] V. MUKHANOV, H. FELDMAN, and R. BRANDENBERGER, ‘Theory of cosmological perturbations’, *Phys. Rep.* **215**, 203, 1992.
- [11] D. LYTH and A. RIOTTO, ‘Particle physics models of inflation and the cosmological density perturbation’, *Phys. Rep.* **314**, 1, 1999.
- [12] A. GUTH and E. WEINBERG, ‘Could the Universe have recovered from a slow first-order phase transition?’, *Nuc. Phys. B* **212**, 321, 1983.
- [13] A. ALBRECHT and P. STEINHARDT, ‘Cosmology for grand unified theories with radiatively induced symmetry breaking’, *Phys. Rev. Lett.* **48**, 1220, 1982.
- [14] D. SALOPEK and J.R. BOND, ‘Nonlinear evolution of long-wavelength metric fluctuations in inflationary models’, *Phys. Rev. D* **42**, 3936, 1990.
- [15] F. LUCCHIN and S. MATARRESE, ‘Power-law inflation’, *Phys. Rev. D* **32**, 1316, 1985.
- [16] D. POLARSKI and A. STAROBINSKY, ‘Semiclassicality and decoherence of cosmological perturbations’, *Class. Quant. Grav.* **13**, 377, 1996.
- [17] A. LICHNEROWICZ, *Théories relativistes de la gravitation and de l'électromagnétisme*,

- Masson, 1955; W. ISRAEL, ‘Singular hypersurfaces and thin shells in general relativity’, *Nuovo Cimento B* **44**, 1, 1966 and correctif dans *Nuovo Cimento B* **48**, 463, 1966; G. DARMOIS, *Les équations de la gravitation einsteinienne*, Mémorial des sciences mathématiques **XXV**, Gauthier-Villars, 1927; K. LANZOS, ‘Flächenhafte Verteilung des Materie in der Einsteinschen Gravitationstheorie’, *Ann. Phys. (Leipzig)* **74**, 518, 1924.
- [18] B. CARTER, ‘Essentials of classical brane dynamics’, *Int. J. Theor. Phys.* **40** (2001) 2099.
 - [19] J. C. HWANG and E.T. VISHNIAC, ‘Gauge-invariant joining conditions for cosmological perturbations’, *Astrophys. J.* **382**, 363, 1991.
 - [20] N. DERUELLE and V. MUKHANOV, ‘On matching conditions for cosmological perturbations’, *Phys. Rev. D* **52**, 5549, 1995.
 - [21] J. MARTIN and D. J. SCHWARZ, ‘The influence of cosmological transitions on the evolution of density perturbations’, *Phys. Rev. D* **57**, 3302, 1998.
 - [22] V. MUKHANOV, *Sov. Phys. JETP* **68**, 1297, 1988; M. SASAKI, *Prog. Theor. Phys.* **76**, 1036, 1986.
 - [23] A. GRIB, S. MAMAEV and V. MOSTEPANENKO, *Quantum effects in strong external fields*, Atomizdat, Moscou, 1980.
 - [24] N. BIRELL and P. DAVIES, *Quantum fields in curved space*, Cambridge University Press, 1982.
 - [25] D. LYTH and E.D. STEWART, ‘A more accurate analytic calculation of the spectrum of cosmological perturbations produced during inflation’, *Phys. Lett. B* **302**, 171, 1993.
 - [26] E.D. STEWART and J.-O. GONG, ‘The density perturbation power spectra to second order corrections in the slow-roll expansion’, *Phys. Lett. B* **510**, 1, 2001.
 - [27] A. KOSOWSKY and M. TURNER, ‘CBR anisotropy and the running of the scalar spectral index’, *Phys. Rev. D* **52**, R1739, 1995.
 - [28] D. SPERGEL *et al.*, ‘First year WMAP observations: determination of the cosmological parameters’, *Astrophys. J. Suppl.* **148**, 175, 2003.
 - [29] E. KOMATSU *et al.*, ‘Five-year WMAP observations: cosmological interpretation’, *Astrophys. J. Suppl.* **180**, 330, 2009.
 - [30] S. LEACH and A. LIDDLE, ‘Constraining slow-roll inflation with WMAP and 2dF’, *Phys. Rev. D* **68**, 123508, 2003.
 - [31] A. LIDDLE and S. LEACH, ‘How long before the end of inflation were observable perturbations produced?’, *Phys. Rev. D* **68**, 103503, 2003; S. DODELSON and L. HUI, ‘A horizon ratio bound for inflationary fluctuations’, *Phys. Rev. Lett.* **91**, 131301, 2003.
 - [32] D. LYTH, ‘What would we learn by detecting a gravitational wave signal in the cosmic microwave background anisotropy?’, *Phys. Rev. Lett.* **78**, 1861, 1997.
 - [33] L. KNOX and M. TURNER, ‘Detectability of tensor perturbations through anisotropy of the cosmic background radiation’, *Phys. Rev. Lett.* **73**, 3347, 1994.
 - [34] L. KNOX and Y.-S. SONG, ‘A limit on the detectability of the energy scale of inflation’, *Phys. Rev. Lett.* **89**, 011303, 2002; M. KESDEN, A. COORAY, and M. KAMIONKOWSKY, ‘Separation of gravitational wave and cosmic-shear contributions to cosmic microwave background polarization’, *Phys. Rev. Lett.* **89**, 011304,

- 2002.
- [35] J. LIDSEY *et al.*, ‘Reconstructing the inflaton potential – an overview’, *Rev. Mod. Phys.* **69**, 373, 1997.
 - [36] M. HOFFMAN and M. TURNER, ‘Kinematic constraints to the key inflationary observables’, *Phys. Rev. D* **64**, 023506, 2001; W. KINNEY, ‘Inflation: flow, fixed points and observables to arbitrary order in slow roll’, *Phys. Rev. D* **66**, 083508, 2002; R. EASTHER and W. KINNEY, ‘Monte Carlo reconstruction of the inflationary potential’, *Phys. Rev. D* **67**, 043511, 2003.
 - [37] A. LIDDLE, ‘On the inflationary flow equations’, *Phys. Rev. D* **68**, 103504, 2003.
 - [38] Y. SHTANOV, J. TRASCHEN and R. BRANDENBERGER, ‘Universe reheating after inflation’ *Phys. Rev. D* **51**, 5438, 1995; L. KOFFMAN, A. LINDE and A. STAROBINSKY, ‘Towards the theory of reheating after inflation’, *Phys. Rev. D* **56**, 3258, 1997; P. GREENE *et al.*, ‘Structure of resonance in preheating after inflation’, *Phys. Rev. D* **56**, 6175, 1997.
 - [39] G. FELDER *et al.*, ‘Dynamics of symmetry breaking and tachyonic preheating’, *Phys. Rev. Lett.* **87**, 011601, 2001.
 - [40] S.Y. KHLEBNIKOV and I.I. TKACHEV, ‘Relic gravitational waves produced after preheating’, *Phys. Rev. D* **56**, 653, 1997.
 - [41] J.-F. DUFAUX *et al.*, ‘Theory and numerics of gravitational waves from preheating after inflation’, *Phys. Rev. D* **76**, 123517, 2007.
 - [42] LATTICEEASY:
<http://physics.stanford.edu/gfelder/latticeeasy/>;
G. Felder and I. Tkachev, *LATTICEEASY: a program for lattice simulations of scalar fields in an expanding Universe*, [hep-ph/0011159].
 - [43] A. VILENKO, ‘Birth of inflationary Universes’, *Phys. Rev. D* **27**, 2848, 1983; A. LINDE, ‘Eternally existing self-reproducing chaotic inflationary Universe’, *Phys. Lett. B* **175**, 395, 1986; A.S. GONCHAROV, A.D. LINDE and V.F. MUKHANOV, ‘The global structure of the inflationary Universe’, *Int. J. Mod. Phys. A* **2**, 561, 1987.
 - [44] A. STAROBINSKY, ‘Stochastic de Sitter (inflationary) stage in the early Universe’, *Lect. Notes Phys.* **246**, 107, 1986; A. STAROBINSKY and J. YOKOYAMA, ‘Equilibrium state of a self-interacting scalar field in the de Sitter background’, *Phys. Rev. D* **50**, 6357, 1994; A. LINDE, D. LINDE and A. MEZHLUMIAN, ‘From the Big-Bang theory to the theory of a stationary Universe’, *Phys. Rev. D* **49**, 1783, 1994.
 - [45] S. WINITZSKI and A. VILENKO, ‘Effective noise in a stochastic description of inflation’, *Phys. Rev. D* **61**, 084008, 2000.
 - [46] J. MARTIN and M. MUSSO, ‘Stochastic quintessence’, *Phys. Rev. D* **71**, 063514, 2005.
 - [47] S. CHANDRASEKHAR, ‘Stochastic problems in physics and astronomy’, *Rev. Mod. Phys.* **15**, 1, 1943.
 - [48] D.S. SALOPEK and J.R. BOND, ‘Stochastic inflation and nonlinear gravity’, *Phys. Rev. D* **43**, 1005, 1991.
 - [49] A. BORDE and A. VILENKO, ‘Eternal inflation and the initial singularity’, *Phys. Rev. Lett.* **72**, 3305, 1993; A. BORDE and A. VILENKO, ‘Singularity in inflationary

- cosmology: a review', *Int. J. Mod. Phys. D* **5**, 813, 1996.
- [50] A.R. LIDDLE, A. MAZUMDAR, and F.E. SCHUNCK, 'Assisted inflation', *Phys. Rev. D* **58**, 061301, 1998.
- [51] S. DIMOPOULOS et al., 'N-inflation', [[hep-th/0507205](#)].
- [52] D. LANGLOIS, 'Correlated adiabatic and isocurvature perturbations from double inflation', *Phys. Rev. D* **59**, 123512, 1999.
- [53] D. POLARSKI and A.A. STAROBINSKY, 'Isocurvature perturbations in multiple inflationary models', *Phys. Rev. D* **50**, 6123, 1994.
- [54] N. BARTOLO, S. MATARRESE and A. RIOTTO, 'Adiabatic and isocurvature perturbations from inflation: power spectra and consistency relations', *Phys. Rev. D* **64**, 123504, 2001.
- [55] A. D. LINDE and V. F. MUKHANOV 'Non-Gaussian isocurvature perturbations from inflation' *Phys. Rev. D* **56**, R535, 1997.
- [56] T. MOROI and T. TAKAHASHI, 'Effects of cosmological moduli fields on cosmic microwave background', *Phys. Lett. B* **522**, 215, 2001.
- [57] D. LYTH and D. WANDS, 'Generating the curvature perturbations without an inflaton', *Phys. Lett. B* **524**, 5, 2002.
- [58] L. KOFMAN, *Probing string theory with modulated cosmological fluctuations*, [[astro-ph/0303614](#)].
- [59] G. DVALI, A. GRUZINOV and M. ZALDARRIAGA, 'A new mechanism for generating density perturbations from inflation', *Phys. Rev. D* **69**, 023505, 2004.
- [60] F. BERNARDEAU, L. KOFMAN and J.-P. UZAN, 'Modulated fluctuations from hybrid inflation', *Phys. Rev. D* **70**, 083004, 2004.
- [61] N. BARTOLO et al., 'Non-Gaussianity from inflation: theory and observations', *Phys. Rep.* **402**, 103, 2004.
- [62] J. MALDACENA, 'Non-Gaussian features of primordial fluctuations in single field inflationary models', *JHEP* **0305**, 013, 2003.
- [63] A. STAROBINSKY, 'Beyond the simplest inflationary cosmological models', *Grav. Cosmol.* **4**, 88, 1998.
- [64] K. NAKAMURA, 'Second-order gauge invariant cosmological perturbation theory', *Prog. Theor. Ph.* **117**, 7, 2007.
- [65] S. WEINBERG, 'Quantum contributions to cosmological correlations', *Phys. Rev. D* **72**, 043514, 2005.
- [66] D. SEERY and J.E. LIDSEY, 'Primordial non-Gaussianities in single field inflation', *JCAP* **0506**, 003, 2005.
- [67] P.J.E. PEEBLES, 'An isocurvature model for early galaxy assembly', *Astrophys. J.* **483**, L1, 1997.
- [68] T.J. ALLEN et al., 'Non-Gaussian density perturbations in inflationary cosmologies', *Phys. Lett. B* **197**, 66, 1987.
- [69] A. GANGUI and S. MOLLERACH, 'Cosmic microwave background non-Gaussian signatures from analytical texture models', *Phys. Rev. D* **54**, 4750, 1996; R. DURRER, M. KUNZ and A. MELCHIORRI, 'Cosmic structure formation with topological defects', *Phys. Rep.* **364**, 2002.
- [70] F. BERNARDEAU and J.-P.UZAN, 'Non-Gaussianity in multifield inflation', *Phys. Rev. D* **66**, 103506, 2002.

- [71] R. BRANDENBERGER and J. MARTIN, ‘The trans-Planckian problem of inflationary cosmology’, *Phys. Rev. D* **63**, 123501, 2001.
- [72] M. LEMOINE *et al.*, ‘The stress-energy tensor for trans-Planckian cosmology’, *Phys. Rev. D* **65**, 023510, 2002.
- [73] A. MESSIAH, *Mécanique quantique*, Dunod, 1964; See also J. MARTIN and D.J. SCHWARZ, ‘WKB approximation for inflationary cosmological perturbations’, *Phys. Rev. D* **67**, 083512, 2003 for the cosmological application of this method.
- [74] J. NIEMEYER and R. PARENTANI, ‘Trans-Planckian dispersion and scale-invariance of inflationary perturbations’, *Phys. Rev. D* **64**, 101301, 2001.
- [75] A. KEMPF and J. NIEMEYER, ‘Perturbation spectrum in inflation with cutoff’, *Phys. Rev. D* **64**, 103501, 2001.
- [76] U. DANIELSSON, ‘A note on inflation and trans-Planckian physics’, *Phys. Rev. D* **66**, 023511, 2002.
- [77] J. MARTIN and C. RINGEVAL, ‘Superimposed oscillations in the WMAP data?’, *Phys. Rev. D* **69**, 064406, 2004.
- [78] V. MUKHANOV, R. ABRAMO and R. BRANDENBERGER, ‘On the backreaction problem for gravitational perturbations’, *Phys. Rev. Lett.* **78**, 1624, 1997; R. ABRAMO, R. BRANDENBERGER and V. MUKHANOV, ‘The energy-momentum tensor for cosmological perturbations’, *Phys. Rev. D* **56**, 3248, 1997.
- [79] T. TANAKA, *A comment on trans-Planckian physics in inflationary Universe*, [[astro-ph/0012431](#)]; A.A. STAROBINSKY, ‘Robustness of inflationary perturbation spectrum to trans-Planckian physics’, *Pisma Zh. Eksp. Teor. Fiz.* **73**, 415, 2001.
- [80] I. TERENO *et al.*, ‘Joint cosmological parameters forecast from CFHTLS-cosmic shear and CMB data’, *Astron. Astrophys.* **429**, 383, 2005.
- [81] J.-P. UZAN, U. KIRCHNER and G.F.R. ELLIS, ‘WMAP data and the curvature of space’, *Month. Not. R. Astron. Soc.* **344**, L65, 2003.
- [82] M. BUCHER, A.S. GOLDHABER and N. TUROK, ‘An open inflation from inflation’, *Phys. Rev. D* **52**, 3314, 1995.
- [83] A. LASENBY and C. DORAN, ‘Closed Universes, de Sitter space and inflation’, *Phys. Rev. D* **71**, 063502, 2005.
- [84] T. PEREIRA, C. PITROU and J.-P. UZAN, ‘Theory of cosmological perturbation in an anisotropic Universe’, *JCAP* **09**, 006, 2007; C. PITROU, T. PEREIRA and J.-P. UZAN, ‘Predictions from an anisotropic inflationary era’, *JCAP* **04**, 004, 2008.

Part III

Beyond the standard models

Grand unification and baryogenesis

The standard model of the electroweak and strong interactions (Chapter 2), combined with Einstein gravity (Chapter 1), provides a representative picture of our knowledge in terms of the fundamental interactions. Nevertheless, as stressed in the two introductory chapters, this description of the fundamental interactions should not be considered as the last word. The purpose of this chapter is therefore to try and go one step further (Grand Unification), and look at the cosmological consequences.

We first discuss the most natural extension that one can bring to the description of the non-gravitational sector, namely Grand Unification (in Section 9.2). This theory aims to describe all the non-gravitational interactions on the same footing, using nothing but the well-controlled techniques developed for quantum field theory. Being well motivated from an experimental point of view, it can really be seen as the simplest possible extension, in the Ockham razor sense, one can look for ‘beyond the standard model’.

Such an extension is expected to apply in the high-energy regime, and, as the energy decreases (or the temperature in a cosmological setup), one should recover the standard model. We thus expect phase transitions, comparable to the electroweak transition discussed in Chapter 2.

The dynamics of the various possible transitions, and particularly those taking place at higher energies, can have important consequences for the evolution of the Universe. In particular, they can induce baryogenesis, i.e. provide a mechanism explaining why the Universe as we observe it is mostly made of baryons and not antibaryons. In Section 9.3, we describe the different mechanisms via which it is possible to produce the observed asymmetry between particles and antiparticles.

9.1 Interactions

In the electroweak model, the coupling constants of the electromagnetic and weak interactions are different not only after the symmetry breaking, but also before, which seems to contradict the very idea of unification. The reason for this fact is that the unification group, in this case $SU(2)_L \times U(1)_Y$, is a product of two different simple groups, and hence is not a simple group itself. To really unify interactions, there should be only one coupling constant; this implies a simple unification gauge group.

9.1.1 Superposition principle

9.1.1.1 Fock space

Quantum theory considers physical states associated with the eigenvalues of the observables. The superposition principle implies that linear combinations of physical states are also physical states of the theory. This is justified as long as the equations of the theory are linear. We can then construct a space of states (one particle Hilbert space, and then a Fock space as the tensor product of Hilbert spaces with increasing numbers of particles), and in particular a basis on which any physical state can be expanded.

Furthermore, the Lagrangians of the field theories described in Chapter 2 introduce different coupling terms between the fields, which translates into interaction terms appearing in the equations of motion.

For a single field, when quantizing, these terms translate into self-interactions of the field, i.e. interactions between the different particles, in other words excitations of this field.

Let us consider the example of a free massive scalar field, ϕ , with potential $V(\phi) = \frac{1}{2}m^2\phi^2$. Its Klein–Gordon equation (2.60) is then linear. So, in terms of particles, any state whose evolution is described by this equation will satisfy the quantum superposition principle. Such a state is thus well defined, remaining an eigenstate of the Hamiltonian at any instant of time. For instance, the one-particle state, $|p\rangle = \hat{a}_p^\dagger |0\rangle$, conserves the same form during its evolution and thus describes a particle with momentum p .

Let us now assume that the potential contains a self-interaction term, for example, $V_i = \frac{1}{4}\lambda\phi^4$. The Klein–Gordon equation then becomes

$$(\square - m^2)\phi = \lambda\phi^3. \quad (9.1)$$

This equation is not linear and is not satisfied by the one-particle states, amongst others. So we cannot hope to have a complete description of these particle states even if they can be used to compute the interaction probabilities to a good approximation. Indeed, even though the particles we observe must interact in order for us to detect them, they are still states with a fixed number of particles.

In practice, the states with a given number of particles are considered as asymptotic states, defined in regions for which the interaction is negligible; usually, this means when the particles one considers are far from each other. Non-gravitational interactions effectively observed in nature act on very short distances and times. Such interactions include, for instance, scatterings that change the physical state, or even decays.

We can thus make the hypothesis that the interaction term is negligible most of the time, so that the particle states in the Fock space, in principle rigorously defined only asymptotically, represent a good approximation to the physical states. At least we assume that they always represent a valid basis for the real Fock space. This is an extremely important hypothesis since it is also in terms of these states that experiments (measurements) are made.

9.1.1.2 Tomonaga–Schwinger equation

In non-relativistic quantum mechanics, time is somewhat different from space as long as one uses a Hamiltonian treatment. Indeed, the evolution equation of the field ϕ is, in most cases, non linear. To obtain a linear equation, and thus to be able to apply the superposition principle, it is necessary to introduce a wavefunctional Ψ , and consider time-independent states that are defined throughout space, i.e. with the functions $\phi(\mathbf{x})$, forming a basis $\{|\phi(\mathbf{x})\rangle\}$, for Ψ . The basic object will then be $\Psi[\phi(\mathbf{x}), t] \equiv \langle \phi(\mathbf{x}) | \Psi(t) \rangle$. Recall what is actually done in ordinary quantum mechanics in which the evolution equations are also non-linear for the position and momentum operators. These non-linearities are avoided by considering the evolution equation of the wavefunction, $\psi(\mathbf{x}, t) = \langle \mathbf{x} | \psi(t) \rangle$. The method is analogous here. In quantum field theory, one can interpret the field as the analogue of the position operator. So, by introducing the functional Ψ , the analogue to the wavefunction ψ of quantum mechanics, we should be able to obtain a linear equation of evolution, independently of the form of the potential.

We recall that the action (1.106) of the field ϕ can be used to define a conjugate momentum π , defined by (2.76), and thus a Hamiltonian that is a functional of both ϕ and π ,

$$\mathcal{H}[\phi(\mathbf{x}), \pi(\mathbf{x})] = \frac{1}{2}\pi^2 + \frac{1}{2}(\nabla\phi)^2 + V(\phi), \quad (9.2)$$

which is a simple generalization of (2.77) for an arbitrary potential.

The evolution equation of the functional $\Psi[\phi(\mathbf{x}), t]$ is obtained by making the correspondence, in the ‘ ϕ ’ representation, in which the field is simply a multiplicative operator, and its momentum conjugate a field derivative according to $\pi(\mathbf{x}) \rightarrow -i\delta/\delta\phi(\mathbf{x})$. The commutation relations (2.20) are obviously satisfied by construction. We can then obtain the associated Schrödinger equation for the functional Ψ , namely

$$\boxed{i\frac{d}{dt}\Psi[\phi(\mathbf{x}), t] = \hat{H}\Psi[\phi(\mathbf{x}), t]} \\ = \int d^3y \left\{ -\frac{\delta^2}{\delta\phi(y)^2} + (\nabla_y\phi)^2 + 2V[\phi(y)] \right\} \Psi[\phi(\mathbf{x}), t]. \quad (9.3)$$

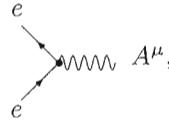
This functional equation is called the Tomonaga–Schwinger equation [1]. In general, it is extremely complicated to solve, which is why we study simple particle-exchange situations, in which case it is sufficient to compute the values of the creation and annihilation operators between states with a fixed number of particles. We notice, furthermore, that (9.3) is linear, which is normal since it is essentially the Schrödinger equation in functional form. Even though we can consider very non-linear interaction terms (ϕ^4 for instance), the superposition principle still applies in field theory with interactions, but for the functional Ψ .

9.1.2 N -particle states

When interactions are taken into account, the one- or many-particle states are no longer eigenstates of the Hamiltonian. This implies, in particular that the number of particles is not necessarily constant in time during the interaction process.

9.1.2.1 Particle exchange

We can interpret interactions as the exchange of virtual particles. This is the case, for example, for electromagnetism for which the interaction between two electrons results from the exchange of a virtual, i.e. unobservable, photon. In the Lagrangian (2.156), one can see that this exchange arises from the term $-2g_2 \sin \theta_w \bar{e} \gamma_\mu A^\mu e$, which is graphically represented in the form



where the vertex, i.e. the central point, has the value $-2g_2 \sin \theta_w$ (see, for example, Refs. [7, 8] of Chapter 2 for more information on the use of Feynman diagrams in particle physics).

The evolution of the functional Ψ , and thus of the state of the system for any $t > t_0$ can be obtained through exponentiation. The functional Schrödinger equation has the same form as the usual Schrödinger equation and can thus be integrated formally in the same way. Thus,

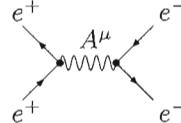
$$\Psi [A^\mu(\mathbf{x}), e(\mathbf{x}), \dots, t] = \exp \left\{ i \hat{H} [A^\mu(\mathbf{x}), e(\mathbf{x}), \dots] (t - t_0) \right\} \Psi \Big|_{t_0}, \quad (9.4)$$

and the state of the system is known at any time, provided that we know it at an initial time t_0 .

Just as in ordinary quantum mechanics, it is possible to shift from the Schrödinger representation to the Heisenberg one in which the states depend on time, but not the operators that act on these states. The evolution of a state is given by

$$|A^\mu(\mathbf{x}), e(\mathbf{x}), \dots, t\rangle = e^{i\hat{H}(t-t_0)} |A^\mu(\mathbf{x}), e(\mathbf{x}), \dots, t_0\rangle. \quad (9.5)$$

Expanding the exponential of equations (9.4) or (9.5), we obtain a series of terms involving higher and higher powers of the interaction. Graphically, this can, for instance, be represented by the exchange of a photon between a positron and an electron. For example, the diagram,



has two vertex points, and is proportional to $4g_2^2 \sin^2 \theta_w$.

9.1.2.2 Transitions

Let us go back to the simpler case of a real self-interacting scalar field, and let us imagine that we perform an experiment to measure the interactions between the excitations of this field. A general asymptotic state is thus a state in the Fock space

defined for free particles, and the initial $|\Psi_{\text{in}}\rangle$ and final $|\Psi_{\text{out}}\rangle$ states then necessarily take the form

$$|\Psi_{\text{in}}\rangle = a_{k_1}^\dagger a_{k_2}^\dagger \cdots a_{k_{N_{\text{in}}}}^\dagger |0\rangle, \quad |\Psi_{\text{out}}\rangle = a_{p_1}^\dagger a_{p_2}^\dagger \cdots a_{p_{N_{\text{out}}}}^\dagger |0\rangle, \quad (9.6)$$

where N_{in} and N_{out} count particles, respectively, entering and exiting the collision.

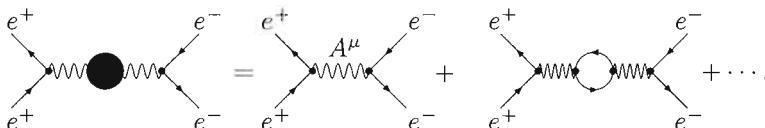
Since the functional Schrödinger equation (9.3) is linear, probability conservation (2.49) implies that the states $|\Psi_{\text{in}}\rangle$ and $|\Psi_{\text{out}}\rangle$ must be related by a unitary transformation

$$|\Psi_{\text{out}}\rangle = U(t_{\text{out}}, t_{\text{in}}) |\Psi_{\text{in}}\rangle, \quad \text{with} \quad U^\dagger(t_{\text{out}}, t_{\text{in}}) = U^{-1}(t_{\text{out}}, t_{\text{in}}). \quad (9.7)$$

This relation describes the transition from the state $|\Psi_{\text{in}}\rangle$ at the time t_{in} to the state $|\Psi_{\text{out}}\rangle$ at the time t_{out} . In the case of states with fixed entering and exiting numbers of particles such as (9.6), this transition operator can be seen as a matrix containing an infinite number of elements that give the transition probabilities. This is the S matrix.

9.1.2.3 Coupling constants

If we are interested in the full series, i.e. summing up over all possible cases, one sees that the complete interaction can be represented diagrammatically in the form



where the dots represent all the terms in the expansion that are not shown explicitly. The first diagram is a *tree* diagram (each part is connected to the others by simple branches), which corresponds to the exchange of a single photon. The second diagram is a *one-loop* diagram and involves the formation of a virtual pair $e^+ - e^-$ that forms and annihilates in the loop while the exchanged photon propagates. This term introduces the coupling constant to the fourth power. As this constant is supposed to be small, one expects that the loop term is merely a correction to the tree diagram. This is even more so when the $e^+ - e^-$ pair is virtual, i.e. the exchanged photon energy need not be higher than twice the mass of the electron for this correction term to be present (classically, this would be simply impossible). However, as the interaction energy between the original electron and the positron grows large, the $e^+ - e^-$ loop turns less and less virtual, and so the term becomes more significant. This implies that this correction in fact depends on the interaction energy. But it also depends on the masses of the particles involved in the loop: one thus readily understands that the variation on the energy of the coupling constants depends on the spectrum of the particles existing in the theory through possible virtual loops (see Chapter 10 for implications of this property).

The total diagram can be interpreted as the effective interaction between an electron and a positron. This amounts to taking into account all the loop corrections by including them in the value of the coupling ‘constant’ of electromagnetism. From this

point of view, we recover the usual electromagnetic interaction but with a coupling constant that now depends on the energy.

It is possible, although beyond the scope of this book, to compute the corrections explicitly and to predict precisely how the coupling constant of the various interactions vary with the energy. For instance, for electromagnetism, we know that $e_0^2 \simeq 1/137$ at low energy, and measure $e_z^2 \simeq 1/128$ when the interaction energy corresponds to the mass of the intermediate boson Z^0 . This measurement, which was made at CERN, is in perfect agreement with theoretical computations (see Ref. [11] of Chapter 2)

9.2 Grand unification

The energy dependence of the coupling constants leads, as will be seen later, to postulate the existence of a unification beyond the one yet realized in the framework of the standard model. Such an extension is also motivated from a theoretical point of view, due to the numerous problems with the standard model.

9.2.1 Problems of the standard model

Despite being compatible with all the experiments available until recently, the standard model of the electroweak and strong interactions suffers from serious theoretical drawbacks and, as was shown in recent years, also from experimental problems. It should be emphasized in particular that the Higgs boson, which is central to its theoretical construction, has still not been detected.

9.2.1.1 The arbitrariness

As presented in Chapter 2, the standard model of particle physics relies on the gauge group $SU(3)_c \times SU(2)_L \times U(1)_Y$. Some might feel uneasy with such a structure, as the reason why this particular group determines the symmetries is unexplained; others might, however, argue that no theory is supposed to explain its own structures. More annoying perhaps, however, is the fact that the model contains three families of quarks and leptons, which looks like an again unexplained replication of the same theory, but at different energy scales. These particles are placed in representations that appear to be completely arbitrary.

The most severe theoretical problem from which the standard model suffers is probably the proliferation of free parameters, which are exclusively determined from experimental measurements and are not computable from first principles. While gravity only introduces one free parameter, the coupling constant G_N , the electroweak and strong theory requires 19 of them! These are:

- the 3 coupling constants associated with the three gauge groups,
- the 9 masses of the massive fermions,
- the 3 angles and the phase of the Cabibbo–Kobayashi–Maskawa matrix (or CKM matrix, described in the following section),
- the vacuum expectation value of the Higgs field $|\langle \phi \rangle| = \eta$, as well as its self-coupling constant λ ,
- and finally the parameter ϑ of the strong interaction.

Taking into account the possibility of a mass for the neutrino, then there are 3 additional masses, one for each generation, and as a consequence new mixing angles and phases. This extraordinarily simple extension raises the number of free parameters to more than 25, which is considered excessive for a theory supposed to explain three of the four fundamental interactions in a unified way.

9.2.1.2 The CKM matrix

Unlike leptons, all the quarks appear in left and right forms so that they are all massive. For the three generations of quarks as well as leptons, it is convenient to generalize (2.166) by use of the substitution

$$\begin{aligned} e &\mapsto e'_A \equiv (e', \mu', \tau') \\ \nu_e &\mapsto \nu'_A \equiv (\nu'_e, \nu'_\mu, \nu'_\tau) \\ u &\mapsto p'_A \equiv (u', c', t') \\ d &\mapsto n'_A \equiv (d', s', b'), \end{aligned} \quad (9.8)$$

where a quantity with a prime indicates an eigenstate of SU(2). The index ‘A’ represents a generation index, and the quarks are designated in a generic way, by analogy with the isospin nomenclature **I** for which the proton (*p*) has the eigenvalue $I_3(p) = +\frac{1}{2}$ and the neutron (*n*) has $I_3(n) = -\frac{1}{2}$. In (9.8), it is replaced by the weak isospin **T**, so we correspondingly would have $T_3(u) = +\frac{1}{2}$ and $T_3(d) = -\frac{1}{2}$. The doublets then become

$$\ell_{AL} = \begin{pmatrix} \nu'_A \\ e'_A \end{pmatrix}_L \quad \text{and} \quad q_{AL} = \begin{pmatrix} p'_A \\ n'_A \end{pmatrix}_L, \quad (9.9)$$

and in a similar way, the singlets also carry a generation index.

The entire standard model can then be simply rewritten by taking the terms already obtained in Chapter 2 and replacing the eigenstates of $SU(2)_L$, commonly called gauge eigenstates or interaction eigenstates, by those of (9.8) and then taking the sum over the generations. We then notice that in general, the coefficients that appear in the (2.166) become matrices

$$\mathcal{L}_{\text{Yukawa}} = \sum_{A,B} - \left[f_{AB}^{(e)} \bar{\ell}_{AL} \cdot \phi e'_{BR} + f_{AB}^{(p)} \bar{q}_{AL} \cdot \tilde{\phi} p'_{BR} + f_{AB}^{(n)} \bar{q}_{AL} \cdot \phi n'_{BR} + \text{h.c.} \right], \quad (9.10)$$

where ϕ is still the Higgs field [$\tilde{\phi}$ is the same as defined below (2.166)].

This substitution implies that for the gauge eigenstates, the mass of the fermions are 3×3 mass matrices (since there are three generations of particles). These matrices are, before any measurement is done, completely arbitrary and in particular they have no reason to be diagonal. In fact, they need not even be symmetric nor Hermitian, and their elements can only be determined experimentally. The mass term of the Lagrangian takes the form

$$\mathcal{L}_{\text{masses}} = \sum_{A,B} -\eta \left[f_{AB}^{(e)} \bar{e}'_{AL} e'_{BR} + f_{AB}^{(p)} \bar{p}'_{AL} p'_{BR} + f_{AB}^{(n)} \bar{n}'_{AL} n'_{BR} + \text{h.c.} \right], \quad (9.11)$$

so that $M_{AB}^{(i)} = \eta f_{AB}^{(i)}$, with $i = e, p$ and n .

The mass matrices $M^{(i)}$ can be diagonalized through bilinear transformations. Notice first of all that for each one of these matrices, the operator MM^\dagger is Hermitian by construction. So it has only positive eigenvalues and can be diagonalized as

$$S^\dagger (MM^\dagger) S = M_d^2, \quad \text{where} \quad M_d^2 = \text{diag } (m_1^2, m_2^2, m_3^2), \quad (9.12)$$

with $m_i^2 \in \mathbb{R}^+$. The mass matrix $M_d \equiv (M_d^2)^{1/2} = \text{diag } (m_1, m_2, m_3)$ obtained in this way is related to M through a bilinear transformation,

$$M_d = S^\dagger M T,$$

where $TT^\dagger = 1$ such that (9.12) is satisfied. The Yukawa interaction term then becomes

$$\bar{\psi}'_L M \psi'_R = \bar{\psi}_L M_d \psi_R,$$

if the mass eigenstates (without the prime) are defined by

$$\bar{\psi}_L \equiv \bar{\psi}'_R S \quad \text{and} \quad \psi_R \equiv T^\dagger \psi'_R.$$

The interaction current between the quarks and the charged gauge bosons is obtained in the same way as for equation (2.156), and we find

$$\mathcal{J}_\mu = \sqrt{2} g_2 \sum_A \bar{p}'_{AL} \gamma_\mu n'_{AL}. \quad (9.13)$$

It can be expressed in terms of the mass eigenstates as

$$\mathcal{J}_\mu = \sqrt{2} g_2 \sum_{A, B, C} \bar{p}_{AL} \left[S_{(p)}^\dagger \right]_{AB} \gamma_\mu \left[S_{(n)} \right]_{BC} n_{CL} = \sum_{A, B} \bar{p}_{AL} \gamma_\mu \left[S_{(p)}^\dagger \cdot S_{(n)} \right]_{AB} n_{BL},$$

so that one can diagonalize the mass terms and write the ones involving the interactions by changing only the quarks d , s and b . The resulting current can then be written as

$$\mathcal{J}_\mu = \sqrt{2} g_2 (\bar{u}, \bar{c}, \bar{t})_L \gamma_\mu V_{CKM} \begin{pmatrix} d \\ s \\ b \end{pmatrix}_L, \quad (9.14)$$

where $V_{CKM} = S_{(p)}^\dagger \cdot S_{(n)}$ is a unitary matrix, called the Cabibbo–Kobayashi–Maskawa matrix.

9.2.1.3 $\hat{\mathcal{CP}}$ violation

For N generations of quarks, the matrix V_{CKM} has N^2 independent matrix elements, once the unitarity condition is taken into account. It is possible to redefine $2N - 1$ of these elements by modifying all but one of the quark phases, since a global rotation of all the phases by the same quantity is always allowed. Moreover, since an orthogonal matrix contains $N(N - 1)/2$ angles, we find that it has $(N - 1)(N - 2)/2$ independent phases. For $N = 2$ generations, this gives a single angle, called the Cabibbo angle and

no phase. For three generations, we obtain three angles and one phase. This phase leads to small, but still measurable violations of the \hat{CP} symmetry.

The current (9.14) indeed couples to the vector boson W^μ such that there exists an interaction term of the form

$$\mathcal{L}_{\text{coupling}} = \mathcal{J}_\mu W^{+\mu} + \text{h.c.}, \quad (9.15)$$

which contains the term

$$\mathcal{L}_{\text{coupling}} \ni V_{ud}\bar{u}\gamma_\mu W^{+\mu}d + V_{ud}^*\bar{d}\gamma_\mu W^{-\mu}u, \quad (9.16)$$

$$\xrightarrow{\hat{CP}} V_{ud}\bar{d}\gamma_\mu W^{-\mu}u + V_{ud}^*\bar{u}\gamma_\mu W^{+\mu}d. \quad (9.17)$$

The second line indicates the transformation of the terms under \hat{CP} , and we have used the fact that $(V^\dagger)_{du} = V_{ud}^*$, where here, for notational simplicity and since there is no risk of confusion, V stands for the CKM matrix with elements V_{ij} between the quarks i and j . This transformation is obtained by generalizing (2.170) and (2.175) (see Ref. [8] of Chapter 2) to the case of the fermions and gauge bosons, taking into account that it is from the action, and not the Lagrangian, that one derives the dynamics. This means in particular that the Lagrangians $\mathcal{L}(x)$ and $\mathcal{L}(-x)$ give exactly the same action $S \propto \int d^3x \mathcal{L}(x) = \int d^3x \mathcal{L}(-x)$. So, they describe the same physical theory that, by a simplifying abuse of notation, we can write as $\mathcal{L}(x) = \mathcal{L}(-x)$.

Equations (9.16) and (9.17) emphasize the fact that if not all the terms of the CKM matrix are real, i.e. if there is a phase that cannot be reabsorbed in the field definition, then the \hat{CP} symmetry is not fulfilled, and processes involving states conjugate to one another under \hat{CP} will not necessarily be produced at the same rate. This is especially important in cosmology since this could provide a way to generate an asymmetry between matter and antimatter.

9.2.1.4 Neutrino oscillations

The same analysis in the leptonic sector does not involve the CKM matrix since neutrinos are supposed to be massless. In particular, this implies that one can postulate, in agreement with experiments, that neutrinos only exist in ‘left-handed’ states, so that there is a unique interaction term between the electrons and neutrinos. Since the mass matrix of the neutrinos vanishes, it is pointless to diagonalize it. We can therefore diagonalize the mass term of the electrons independently of the coupling term. Then, one can simplify the form of the couplings by redefining the neutrino fields, a harmless (unphysical) operation. Of course, if neutrinos turn out to be massive, all of these conclusions will be modified, and we then recover the previous CKM situation, with a few subtle differences, due to the fact that these particles are not subject to the strong interaction.

One way to determine whether the neutrino is truly massless or if its mass is simply too small to have been detected is to look at the oscillations between the different generations. We consider in what follows the simplified case of two mixed species, with a unique mixing angle and no phase. It is not difficult, and even recommended as an exercise, to generalize this calculation to three neutrino species.

Let us assume, for example, that there exists a mixing between the two families of neutrinos ν_e and ν_μ . These two species can be decomposed into a basis of the mass eigenstates ν_1 and ν_2 , of respective masses m_1 and m_2 . If the mixing occurs via the Cabibbo angle, α , we can express the different states by the relations

$$\begin{pmatrix} |\nu_e\rangle \\ |\nu_\mu\rangle \end{pmatrix} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} |\nu_1\rangle \\ |\nu_2\rangle \end{pmatrix} \iff \begin{pmatrix} |\nu_1\rangle \\ |\nu_2\rangle \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} |\nu_e\rangle \\ |\nu_\mu\rangle \end{pmatrix}, \quad (9.18)$$

where the states $|\nu_i\rangle$, $i = 1, 2$ are eigenstates of the Hamiltonian

$$H |\nu_i\rangle = E_i |\nu_i\rangle = \sqrt{\mathbf{p}^2 + m_i^2} |\nu_i\rangle \quad (9.19)$$

for states of momentum \mathbf{p} .

Let us now consider a neutrino produced by a nuclear reaction in the Sun. Since it results from an interaction, it is necessarily a neutrino ν_e or ν_μ in our simplified theory that describes only two interaction states. For concreteness, let us assume that it is an electron neutrino. In the Schrödinger representation, the state $|\psi(t)\rangle$ evolves as (2.50), which gives, with $|\psi(0)\rangle = |\nu_e\rangle$,

$$\begin{aligned} |\psi(t)\rangle &= e^{-iHt} |\nu_e\rangle \\ &= e^{-iHt} (\cos\alpha |\nu_1\rangle + \sin\alpha |\nu_2\rangle) \\ &= e^{-iE_1 t} \cos\alpha |\nu_1\rangle + e^{-iE_2 t} \sin\alpha |\nu_2\rangle \\ &= (\cos^2\alpha e^{-iE_1 t} + \sin^2\alpha e^{-iE_2 t}) |\nu_e\rangle \\ &\quad + \sin\alpha \cos\alpha (e^{-iE_2 t} - e^{-iE_1 t}) |\nu_\mu\rangle. \end{aligned} \quad (9.20)$$

So, the neutrino does not remain constantly in the electron neutrino state, it oscillates between the electronic and muonic states.

From the law (2.43), the probability to detect the neutrino in its electronic state is then

$$\mathcal{P}_\psi(\nu_e) = |\langle \nu_e | \psi(t) \rangle|^2 = |\cos^2\alpha e^{-iE_1 t} + \sin^2\alpha e^{-iE_2 t}|^2, \quad (9.21)$$

where we have used the fact that the basis $\{|\nu_e\rangle, |\nu_\mu\rangle\}$ is orthonormal and assumed an ideal detector, with no loss (all electron neutrinos are detected).

For relativistic neutrinos with $m_i^2 \ll \mathbf{p}^2$, we can expand the energy as $E_i \simeq |\mathbf{p}| + \frac{1}{2}m_i^2/|\mathbf{p}|$, and we find

$$\mathcal{P}_\psi(\nu_e) \simeq 1 - \sin^2 2\alpha \sin^2 \left(\frac{\Delta m^2 L}{4|\mathbf{p}|} \right). \quad (9.22)$$

The distance $L = t$ is that between the Sun and the detector (since neutrinos are relativistic, they essentially travel at the speed of light), and $\Delta m^2 \equiv |m_2^2 - m_1^2|$.

Relation (9.22) indicates that a deficit in the number of electron neutrinos detected compared to that emitted is only possible if neutrinos have a non-vanishing mass and if, moreover, these masses are different for each generation. Knowledge from the nuclear reactions in the Sun makes it possible to predict precisely its rate of neutrinos emission. It has been noticed for some time that the flux of electron neutrinos received is much

lower than that predicted, by a proportion reaching 75% in some experiments. This solar neutrino problem has therefore given the first indication for the possibility that neutrinos are massive.

Finally, we notice that the propagation of neutrinos in matter is different from that in the vacuum since two additional effects contribute to the oscillations. Every active neutrino interacts with quarks and leptons through the exchange of a Z^0 boson, while sterile neutrinos, discussed in what follows, by definition do not undergo this interaction. Moreover, only electron neutrinos and antineutrinos interact with electrons, which compose an important part of matter, through the exchange of W^\pm . Both effects can easily be taken into account by introducing a typical oscillation length, independent of the neutrino energy, but a function of the matter density. It turns out to be of the order of 14 km in the Earth, and around 200 km in the core of the Sun [2].

9.2.1.5 Super-Kamiokande and neutrino masses

Since 1998, the experimental situation has changed. The Super-Kamiokande experiment [3], a neutrino detector,¹ has shown the existence of an angular dependence in the distribution of the electron neutrinos received on Earth. The flux of neutrinos from different zenithal directions is different, which indicates that the number of incident neutrinos is a function of the distance they travelled through the Earth, as expected by the oscillation model.

More recently, the KamLAND experiment [4] in Japan has established the neutrino oscillations even more sharply, as can be seen in Fig. 9.1 where the ratio between the expected distribution of neutrinos without oscillations and the measurements are reported. This experiment consists of a 13-m diameter transparent sphere containing 1000 tonnes of a very pure liquid scintillator and surrounded by 1879 photo-multipliers filling the internal surface of an 18-m diameter container. Electron antineutrinos are detected through the inverse β decay, $\bar{\nu}_e + p \rightarrow e^+ + n$, with an energy threshold of 1.8 MeV. Everything is surrounded by a water Čerenkov detector of 3200 tonnes that absorbs photons and neutrinos from the environment and detects cosmic muons.

Around the KamLAND experiment site, there are 53 electric reactors, with known time-dependent fluxes of neutrino emissions, which allows for a good calibration and elimination of systematic effects. After collecting data between March 2002 and January 2004, the experiment concluded on the existence of oscillations with parameters

$$\Delta m^2 = 7.9_{-0.5}^{+0.6} \times 10^{-5} \text{ eV}^2 \quad \text{and} \quad \tan^2 \alpha = 0.4_{-0.07}^{+0.1},$$

where the mass difference is between the electron neutrino and another, yet unspecified (presumably belonging to one of the two other flavours available in the standard model) species of neutrinos, the analysis being restricted to two species. These numbers have been established by taking into account other experiments dedicated to the observation of solar neutrinos only.

¹Originally conceived to detect proton decay, this large (50 000 tons of water) Čerenkov reservoir, buried deep in the ground in order to shield as much as possible from cosmic rays, turned out, because of both its size and location, to be a very efficient neutrino detector.

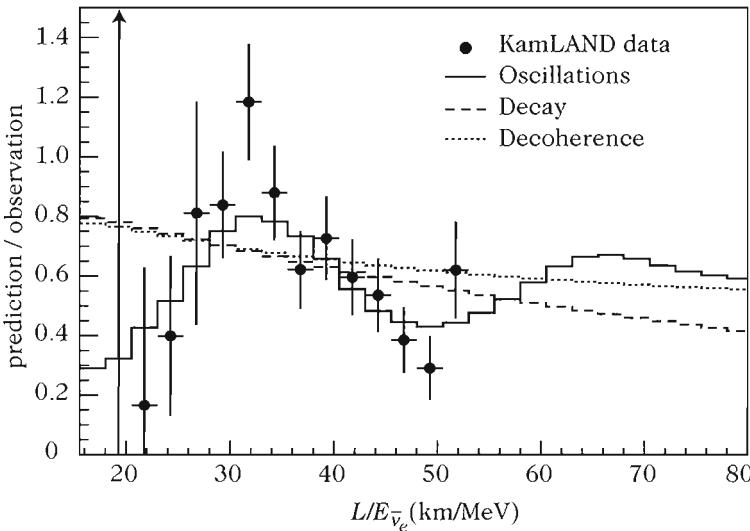


Fig. 9.1 Observed distribution of neutrinos compared to the expected distribution in the absence of oscillations in the KamLAND experiment. The measured points are in disagreement with any possible models (decay or decoherence) other than oscillations.

9.2.2 Unification of the coupling constants

The existence of couplings between the particles implies that the different constants introduced in the total interaction Lagrangian must be renormalized for each value of the energy at which a given interaction happens. The running of the coupling constant with energy is illustrated in Fig. 9.2 for the three non-gravitational interactions, extrapolated from experimental data up to the Z^0 peak (LEP experiment at CERN) and from the properties of the known particles. We notice a very clear tendency of the three curves to converge, indicating the possibility that at very high energy, in practice of the order of 10^{14} GeV, the three coupling constants, which are very different at low energy, are of the same order of magnitude for such high energies, and one can postulate that they could in fact merge to a common value. In this case, the three interactions would be indiscernible since not only would they be of the same intensity, but they also would act in the same way. We can thus assume that the three symmetry groups, colour $SU(3)_c$ from the strong interaction, and $SU(2)_L \times U(1)_Y$ from the electroweak interactions, are gathered into a single group \mathcal{G} that contains them all, a grand unification group. In other words, these three interactions would only be different low-energy manifestations of the same interaction.

It is interesting to note at this level, that the electric charges of the quarks and leptons lead to the (at this stage unexplainable) relation

$$\sum_{\text{particles}} Q_i = 0,$$

where the Q_i are the particle charges: this is achieved if each quark is counted three

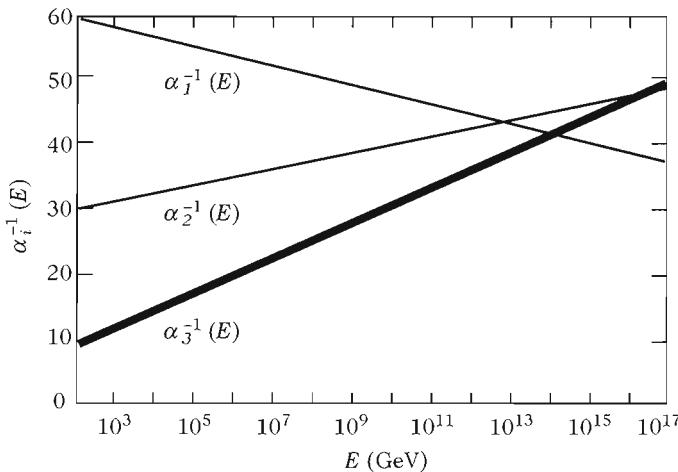


Fig. 9.2 Running of the inverse coupling constants of the standard model of the electroweak (α_1^{-1} and α_2^{-1}) and strong (α_3^{-1}) interactions with interaction energy E . We see that α_1 and α_2 are increasing functions of the energy, whereas α_3 decreases with E . These curves are computed with all the experimental data (average of the values obtained in 1991), and their width reflects the experimental uncertainty, which is more significant for the strong interactions (α_3). The extrapolation of these curves to very high energies indicates a tendency to converge at around 10^{14} – 10^{15} GeV, which we call the grand unified energy. Above this energy, the three interactions have a comparable magnitude and we assume that the three forces are unified, with a symmetry being that of a simple group \mathcal{G} containing $SU(3)_c \times SU(2)_L \times U(1)_Y$.

times to compensate for the fact that quarks have fractional charges. Provided the strong interaction is taken into account, a necessary condition for a unifying theory, this counting is very natural since each strongly interacting particle comes in three colour states.

It so happens that this otherwise incongruous relation is essential in order for the theory to have no anomalies, i.e. to be renormalizable, and thus armed with a sufficient predictive power to make it a physical theory. Even though the electroweak and strong interactions are independent, we notice that the electroweak model is only renormalizable provided that the degrees of freedom arising from chromodynamics are considered. This reinforces the idea of unification. In the framework of such a unification, where we try to describe quarks and leptons in the same multiplets, the electric charge becomes a generator of the unified Lie symmetry group. It is no longer an arbitrary quantity as in the standard model, but derived from the group actions. As such, the electromagnetic charge operator ought to be traceless when acting on a multiplet. The previous relation, which is completely arbitrary (and thus a strange coincidence) in the framework of the usual standard model, becomes a mere consequence of any unification scheme. This is, from our point of view, the most compelling theoretical argument in favour of unification.

9.2.3 Unification models

Historically, it was H. Georgi and S. L. Glashow, in 1974 [5], who proposed the first Grand Unified Theory (GUT), based on the group $SU(5)$. The electroweak theory, introduced in Chapter 2, only represents a partial unification since the constants g_1 and g_2 of the groups $U(1)_Y$ and $SU(2)_L$ are not equal before the symmetry breaking. Technically, this is due to the fact that the unification group is a product of two different simple groups. To achieve a unification with a single coupling constant, the symmetry group should be of the form \mathcal{G}^n , where \mathcal{G} is simple and n an arbitrary integer. For $n \neq 1$, a discrete symmetry should also be imposed between the representations used for \mathcal{G} . It is thanks to this symmetry that the groups can be seen as equivalent, which allows for the presence a single coupling constant despite the presence of several groups.

As seen in Chapter 2, all the simple groups are known and classified. In particular, we know the groups that enjoy complex representations and thus for which it is easier to fit all the known particles since they are already fit into complex representations of $SU(3)_c \times SU(2)_L \times U(1)_Y$. The interesting groups from this point of view are thus $SU(n)$ for $n > 2$, E_6 and $SO(4n+2)$.

9.2.3.1 Pati–Salam model

Before the possibility of a total unification was realized, Pati and Salam [6] proposed, as a first step, to consider each generation of leptons as a fourth colour, thus substituting the group $SU(3)_c$ by the group $SU(4)_c$. In order for this to be possible, the quark and lepton doublets and singlets must have the same properties, and to start with, they must be divided in a similar way. We then require that the doublet pair of $SU(2)_L$, (\mathbf{q}, ℓ) , transforms as the representation $\mathbf{4}$ of $SU(4)_c$, but the set composed of the conjugate particles (\mathbf{q}^c, ℓ^c) should then also be a doublet. Since the neutrino is assumed massless and as a result has only a left-handed component, this requires the introduction of a new so-called ‘sterile’ neutrino state ν_s , bearing no quantum number of the standard model. This neutrino completes the doublet (\mathbf{q}^c, ℓ^c) , with the identification $\nu^c = \nu_s$. The set composed of the right electron and the sterile neutrino is then a doublet of $SU(2)_R$, such that the symmetry of the Pati–Salam model is $G_{PS} \sim SU(4)_c \times SU(2)_L \times SU(2)_R$.

The group $SU(4)_c$ contains the group $SU(3)_c$ of ordinary colours as a subgroup, but also a $U(1)$ which should be identified with $B - L$, where B is the baryon number and L is the lepton number. These numbers simply count how many baryons and leptons are in a given state, allowing us to identify these particles in the doublets. The measured properties of the different particles are reproduced as long as the Gell-Mann–Nishijima relation (2.153) is modified in the form

$$Q = T_L^3 + T_R^3 + \frac{1}{2} (B - L). \quad (9.23)$$

We see that we can achieve here a first form of unification: if an $L \leftrightarrow R$ symmetry is imposed, we will only have a single coupling constant for the two groups $SU(2)_L$ and $SU(2)_R$. The Pati–Salam group is present in many symmetry breaking schemes of higher-rank unification groups. The sterile neutrino is then often identified with

the right-handed neutrino. This, in turns, implies the neutrino is massive ... and thus oscillating, in agreement with the data previously discussed.

The Pati–Salam unification model is not completely satisfying from a unification point of view since even though it unifies quarks and leptons, i.e. the matter particles, in an equivalent set, it gives no predictions for the exchange-particle interactions. In other words, it only suggests a partial unification of the particles and not of the interactions. The aim of grand unified theories is precisely to achieve such a unification.

9.2.3.2 $SU(5)$ and beyond

The unification of Pati–Salam is not yet completely satisfying as it still contains two coupling constants. Knowing the unification scale, E_{GUT} , and the way the coupling constants of the standard model evolve till this point, then we have three relations giving g_1 , g_2 and g_3 as a function of the coupling constant of $SU(4)_c$, of that of the two $SU(2)$, and finally of E_{GUT} . These three relations for three constants allow for no prediction, and the three coupling constants can still be seen as arbitrary.

To go further and have only one coupling constant, we should construct a simple group containing the standard model. In this case, the three constants g_1 , g_2 and g_3 will depend on the grand unified constant g_{GUT} and on the unification scale E_{GUT} , making it possible to determine one of the three g_i as a function of the two others. In general, we choose to compute the weak angle (2.157) as a function of the other parameters.

Table 2.2 indicates that the rank of the standard model symmetry group is 4. So, in order for it to be a subgroup of the required unification group, the rank of the latter should be at least equal to 4. Notice, by the way, that since the rank of G_{PS} is equal to 5, it will only appear as a subgroup in the possible schemes if the unification uses a group of rank at least equal to 5.

Still, from Table 2.2, we see that the first simple group of rank ≥ 4 is $SU(5)$. This is actually why it was historically the first candidate to realize grand unification. This group, due to its representations, has another immediate advantage: we can place all the particles, quarks and leptons, in the representations **5** and **10**. This is realized in the following way:

$$\mathbf{5} \rightarrow \begin{pmatrix} \bar{d}_1 \\ \bar{d}_2 \\ \bar{d}_3 \\ e^- \\ \nu \end{pmatrix} \quad \text{and} \quad \mathbf{10} \rightarrow \begin{pmatrix} 0 & \bar{u}_2 & -\bar{u}_1 & u_1 & d_1 \\ -\bar{u}_2 & 0 & \bar{u}_3 & u_2 & d_2 \\ \bar{u}_1 & -\bar{u}_3 & 0 & u_3 & d_3 \\ -u_1 & -u_2 & -u_3 & 0 & e^+ \\ -d_1 & -d_2 & -d_3 & -e^+ & 0 \end{pmatrix}, \quad (9.24)$$

where the representation **10** is an antisymmetric matrix. The quark indices are colour indices of $SU(3)_c$. In (9.24), all the particles are left bi-spinors.

Despite being the most economical model we can think of from a unification group point of view, it is yet not completely satisfying for two reasons, an experimental and a theoretical one. First, it leads to too short a proton lifetime compared to the known constraints (see Section 9.2.4); in its simplest version, $SU(5)$ is simply experimentally ruled out.

From the point of view of the theory itself, the model has a feature that, subjectively speaking, is unpleasant: the aim of unification is not only to describe all the interactions as a single one, but one would also wish to be able to represent all the particles on an equal footing, i.e. into a single rather than into two representations. To do this, we need a group having a 15-dimensional (at least) representation. We then go a bit further and resort to the next possible unification group in the list, namely SO(10). As it so happens, SO(10) has a representation of dimension **16**, in which one can easily fit all the known particles ... plus one! This additional element of the **16** representation of SO(10), is a bi-spinor like the other elements, and must be interpreted as a right-handed neutrino state. So, a general prediction of grand unification beyond SU(5) is the existence of a mass for the neutrino.

All the other unification groups are possible, as long as they satisfy some constraints, such as, for instance, the existence of complex representations [7]. However, for high-rank groups, it becomes necessary, in order to obtain the standard electroweak model at low energy, to rely on numerous Higgs bosons to break the symmetries, which introduce a large number of arbitrary parameters. This reduces considerably the advantage and the ‘natural’ aspect of such models. This is why, in general, we take to be realistic only groups up to rank 6, such as E₆ or SU(7), or of rank 7, such as SO(14) or SU(8). We also find occasionally (and stemming from string theory) the groups E₈ and SU(9), of rank 8, but the profusion of generators, already at this stage, makes the calculations often inextricable.

9.2.3.3 Symmetry-breaking scheme of grand unification

To construct the standard model of particle physics starting from the symmetry of a given grand unification \mathcal{G} , Higgs fields should be introduced to reduce the symmetry progressively until reaching that of the standard model, namely $SU(3)_c \times SU(2)_L \times U(1)_Y$. The way we choose to distribute these scalar fields in the different possible representations of \mathcal{G} indicates what will be the symmetry-breaking scheme. For instance, if we choose the representation of dimension **126** of SO(10) then, when the Higgs field takes a non-vanishing expectation value, the symmetry is reduced to $SU(5) \times Z_2$.

To understand this mechanism it is better to consider a simple example, even though it will not be realistic from a grand unification point of view; this is what we call a ‘toy model’, i.e. a theoretical model having the set of properties of a real theory (at least we hope) but simple enough to be able to ‘play’ with it.

So let us assume that the initial symmetry of the theory is that of the group SU(2). The theory describes the dynamics of a Higgs field ϕ in a representation of SU(2) as well as the gauge bosons A_μ^a associated with the transformations of SU(2). The Lagrangian we are interested in has therefore all the usual kinetic terms already described, as well as a symmetry-breaking potential with a minimum at $\phi^2 = \eta^2$. Depending on the chosen representation for the Higgs field, we will have different breaking schemes, leading to different physical theories at low energy.

For example, if the Higgs field is in the real representation **3** [fundamental representation of SO(3), which is isomorphic to SU(2)], i.e. a vector composed of three scalar fields

$$\phi = \begin{bmatrix} \phi_1(x, t) \\ \phi_2(x, t) \\ \phi_3(x, t) \end{bmatrix},$$

then its squared modulus is simply defined by $\phi^2 \equiv \phi_1^2 + \phi_2^2 + \phi_3^2$, such that there still exists an infinity of possible configurations, distributed on the sphere of the internal space of the Higgs field, corresponding to the symmetry $SU(2)$. In order to be able to actually perform calculations, we must choose one of these configurations. At this stage, this is completely arbitrary and with no local physical implications². Let us assume then that the vacuum configuration is given by $\phi_3 \neq 0$, and thus

$$\phi_0 = \begin{pmatrix} 0 \\ 0 \\ \eta \end{pmatrix}.$$

The generators of $SU(2)$ expressed in the representation **3** are simply the usual matrices generating the infinitesimal rotations

$$T_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} \quad T_2 = \begin{pmatrix} 0 & 0 & i \\ 0 & 0 & 0 \\ -i & 0 & 0 \end{pmatrix} \quad \text{and} \quad T_3 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

which correspond, respectively, to the rotations around the axis 1, 2 and 3. It is clear that the third generator, T_3 , is annihilated by the vacuum ϕ_0 , i.e. $T_3\phi_0 = 0$. The vacuum is thus still invariant under transformations of the form $\exp(-iaT_3)$, which form a $U(1)$ subgroup of the original $SU(2)$ group. This scheme is thus symbolically expressed in the form $SU(2) \xrightarrow{3} U(1)$. Notice that a different choice of vacuum would have produced another invariant generator, and finally the same scheme, up to a rotation.

Let us now consider the case where ϕ transforms as the complex representation **2**, i.e. it is a doublet

$$\phi = \begin{bmatrix} \phi_1 = \Re(\phi_1) + i\Im(\phi_1) \\ \phi_2 = \Re(\phi_2) + i\Im(\phi_2) \end{bmatrix}.$$

Once the symmetry is broken, we have $|\phi_1|^2 + |\phi_2|^2 = \eta^2$, which is obtained, for instance, by the choice of the vacuum

$$\phi_0 = \begin{pmatrix} 0 \\ \eta \end{pmatrix}.$$

Given the generators of $SU(2)$ of this representation, here the Pauli matrices, it is easy to convince ourselves that no combination of these generators annihilates the vacuum, and as a consequence, no symmetry is left invariant. We thus have the scheme $SU(2) \xrightarrow{2} \{\text{Id}\}$, the group reduces to the identity element.

Finally, a last interesting case relates to multiple breakings, which are often found in realistic theories of grand unification since it is not always possible to break directly

²We shall see later, in Chapter 11, that from a global point of view, the choice is not necessarily harmless as it might lead to the formation of topological defects.

the initial symmetry into that of the standard model. Let us consider again the example of the ϕ in the representation **3** of $SU(3)$, and let us add to this theory another scalar field, χ , which also transforms as **3**. In order to describe completely the dynamics of this new theory, we also introduce a Higgs potential for this new field and a coupling term, assumed to be renormalizable, given, respectively, by

$$V_\chi(\chi) = \frac{\lambda_\chi}{4} (\chi^2 - \eta_\chi)^2 \quad \text{and} \quad V_{\text{coupl}}(\phi, \chi) := \alpha (\phi \cdot \chi)^2,$$

where λ_χ , η_χ and α are constant. To minimize this new global potential, not only do we need $\phi^2 = \eta^2$ and $\chi^2 = \eta_\chi^2$, but also $\phi \cdot \chi = 0$, which implies, once the direction of ϕ is

determined in the same way as in the previous example, that we can take $\chi = \begin{pmatrix} \eta_\chi \\ 0 \\ 0 \end{pmatrix}$

or $\chi = \begin{pmatrix} 0 \\ \eta_\chi \\ 0 \end{pmatrix}$. In both cases, no more symmetry remains, and we therefore have the

multiple scheme $SU(2) \xrightarrow{3} U(1) \xrightarrow{3} \{\text{Id}\}$. Given this series of examples, we reach the conclusion that there can be different ways to reach the same final symmetry, starting from a given symmetry.

With a given initial symmetry, the resulting low-energy physics will be completely determined by the representations according to which the Higgs fields will transform, and their interaction potential; this opens up so many possibilities that external requirements must be imposed for an actual theory to be written down.

All of these considerations indeed directly apply to more realistic groups, and allow us to sketch the main ideas of what could have happened in the primordial Universe from the period where, as we think, grand unification was realized.

9.2.4 Consequences of grand unification

The consequences of grand unification are twofold. At low energy, the unification can manifest itself by small effects in the interaction between particles. It shows up in particular when one studies the proton stability, and might also manifest in interactions involving neutrinos. Furthermore, the existence of a grand unified phase at very high energies, i.e. high temperatures, can have influenced the early evolution of the Universe. Finally, these two kinds of consequences can be mixed, producing effects inherent to particle physics, but whose main influence is sensible during the first phases of the Universe.

9.2.4.1 New interactions

Although experimentally disfavoured, let us, however, reconsider, as a toy model, the prototypical example of GUT theory, based on the group $SU(5)$. An arbitrary transformation of this group is represented in the general form

$$\mathbf{T}_5 = \begin{bmatrix} SU(3)_c & (X, Y) \\ (X, Y) & SU(2)_L \times U(1)_Y \end{bmatrix}, \quad (9.25)$$

when taking into account the interactions of the standard model. If we form a scalar from this matrix with the fields in the representation 5 of equation (9.24), for example, of the form $\bar{\mathbf{5}}\mathbf{T}_5\mathbf{5}$, we immediately see that the non-diagonal terms, i.e. those denoted by X and Y in (9.25), mix the quarks and leptons. In other words, they induce a direct interaction between these particles.

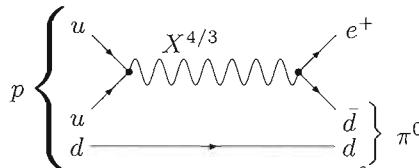
Since quarks have fractional charges, it is easy to convince ourselves that the intermediate X and Y bosons are also fractionally charged, and we have, in our representations,

$$Q_{\text{elec}}(X^\mu) = \pm \frac{4}{3} \quad \text{and} \quad Q_{\text{elec}}(Y^\mu) = \pm \frac{1}{3}.$$

These particles must also have a mass of the order of the grand unification scale, that is, from Fig. 9.2, $M_{X,Y} \simeq 10^{14}$ GeV.

9.2.4.2 Proton decay

The mass obtained above is, however, very tightly constrained by the proton lifetime. Indeed, the above-mentioned interactions also imply that the proton can no longer be stable since diagrams such as



through which a proton decays into a positron and a π^0 , are also possible. The decay width $\Gamma(p \rightarrow \pi^0 e^+)$ for this process is easily computable following the principles of basic field theory and we find

$$\Gamma(p \rightarrow \pi^0 e^+) \equiv \tau_p^{-1} = g_5^4 \frac{m_p^5}{M_X^4}, \quad (9.26)$$

where m_p is the mass of the proton and g_5 the coupling constant of the group SU(5). This constant, also to be read from Fig. 9.2, is somewhat constrained and therefore cannot be made as small as desired. As a result, the lifetime τ_p of the proton is essentially controlled by the mass of the intermediate boson X . In order for the proton to be stable enough, we see that the characteristic energy of grand unification must be quite large. From this constraint and the measured energy dependence of the coupling constants, we see that the grand unified theories are not purely speculative but are also falsifiable (not all speculative theories share this property!). All the current data converge towards a lifetime greater than 10^{29} years in the invisible decay modes (it is therefore an absolute constraint and independent of any theoretical model) and 10^{33} years in the particular mode discussed earlier. Given the expected value of the coupling constant from Fig. 9.2 as well as the energy scale M_X , these numbers are barely compatible with the simplest unification theories; as already mentioned, SU(5) is already ruled out because of this. This is perhaps the strongest theoretical argument in favour of supersymmetric extensions to standard theories.

9.2.5 Neutrino mass

Apart from the SU(5) model, all grand unified theories require the existence of a right-handed neutrino, and thus demand that neutrinos be massive. Besides, the masses measured through the neutrino oscillation experiments are extremely weak, of the order of a thousandth of an electron-volt. This poses, already in the framework of the standard model, a problem of parameter tuning. This problem would appear to be much more severe in the context of grand unification since the energies involved are considerably higher. This is actually not the case, and on the contrary it is even possible to use these high energies to naturally produce very small neutrino masses. This is the ‘see-saw’ mechanism, which gives a very large mass to the right-handed neutrino, thus explaining why it has never been directly produced and observed, and at the same time, by extension, a very small mass for the left-handed neutrino, thus explaining the latter.

9.2.5.1 Dirac and Majorana neutrinos

In general, there are two kinds of mass terms for the neutrinos, the ones for which the neutrino is its own antiparticle, and the others. In the first case, we call it a Majorana neutrino, and it can be completely described by a bi-spinor. The second one requires a true quadri-spinor, and is a Dirac neutrino. The standard model of particle physics has three left-handed neutrinos that are described by bi-spinors ν_{ℓ_L} , with $L \in (e, \mu, \tau)$ and the grand unified models predict the existence of other similar states, say N , generically designed as sterile neutrinos, which we denote as ν_{α_R} , $\alpha = 1, \dots, N$.

Gathering the neutrinos in the multiplet $\nu = \begin{bmatrix} \nu_{\ell_L} \\ (\nu_{\alpha_R})^c \end{bmatrix}$, where the exponent ‘ c ’ stands for the charge conjugation, then the most general mass term that we can write is of the form

$$\mathcal{L}_{\text{mass } \nu} = \frac{1}{2} \bar{\nu}^c \mathbf{M} \nu + \text{h.c.}, \quad (9.27)$$

where the mass matrix \mathbf{M} is decomposed as

$$\begin{bmatrix} 0 & \mathbf{M}_D \\ -(\mathbf{M}_M)^T & \mathbf{M}_M \end{bmatrix}, \quad (9.28)$$

with \mathbf{M}_D and \mathbf{M}_M , respectively, Dirac and Majorana mass matrices. In the case where the eigenvalues of \mathbf{M}_M are very large compared to that of \mathbf{M}_D , we are in the ‘see-saw’ regime, and the eigenstates of \mathbf{M} have effective masses that can be very different.

9.2.5.2 ‘See-saw’ mechanism

Let us illustrate this mechanism in the example of a single generation of leptons and a single sterile neutrino. In this case, we can ignore the indices and the Dirac and Majorana mass matrices become simple numbers. The diagonalization of (9.28) leads to the two eigenvalues

$$M_{\pm} = \frac{1}{2} \left(M_M \pm \sqrt{M_M^2 - 4M_D^2} \right) \sim \begin{cases} M_M - \frac{M_D^2}{M_M}, \\ \frac{M_D^2}{M_M}, \end{cases} \quad (9.29)$$

where we have explicitly considered the case $M_D \ll M_M$. This relation shows that with a mass scale characteristic of grand unification, say $M_M \sim 10^{15}$ GeV ($\sim M_+$), we can obtain a very small mass for the left-handed neutrino (M_-), namely of the order of the meV say, provided that the Dirac masses are around a hundred GeV, i.e. a mass comparable to that of the other particles from the standard model. We also see that the right-handed neutrino naturally keeps a very high mass M_+ , of the order of the unification scale, which explains why it has never been observed.

9.3 Baryogenesis

At very high temperatures there are, even in the standard model, interactions that violate the conservation of the baryon number. In thermal equilibrium, the production of particles must be equal to that of antiparticles so that the total baryon number, i.e. the number of baryons minus the number of antibaryons, must vanish. Observations indicate, however, a strong asymmetry between baryons and antibaryons, at least locally. Furthermore, models containing as many baryons as antibaryons, but distributed in a very inhomogeneous way (to explain why we happen to live in a region highly dominated by baryons only), are very constrained. For instance, in such an inhomogeneous Universe, there must exist frontiers between matter-dominated and antimatter dominated domains. These would have the structure of domain walls and would be regions of contact and thus of annihilation. This means they would emit very intensely in X and γ rays and would be easily detectable. No such region having ever been detected, we tend to believe that there exists an initial asymmetry between particles and antiparticles.

Unless we impose by hand such an initial asymmetry merely to reproduce the data, there must have been some dynamical mechanism at work through which the number of baryons became much larger than that of antibaryons: this mechanism is called baryogenesis. Some theories achieve this at low energies, and in these cases we expect them to be testable in accelerators (in particular, see the BABAR experiment and the tests of the $\hat{C}\hat{P}$ violation in the system $B^0 - \bar{B}^0$ [8]), whereas in other theories, the baryon number is produced at high energy and is then protected from thermalization by the $B - L$ symmetry [9].

9.3.1 Sakharov conditions for baryogenesis

In 1967, Sakharov [10] formulated the conditions under which it is possible to produce an asymmetry between baryons and antibaryons in the Universe. Three ingredients are required:

1. Violation of baryon number. This is of course a necessary condition if we assume that the initial condition is that of a state containing as many baryons as antibaryons, or even containing none of each.

2. \hat{CP} violation. Again, this is a *sine qua non* condition since it implies that the antiparticles do not behave exactly as the particles. If this were not the case, then for each production of a baryon from a particle, there would be production of an antibaryon from the antiparticle with the same probability, and the total number would thus be conserved.
3. Out-of-equilibrium state. This last condition ensures that the production of an arbitrary type of particle via one mechanism is not immediately compensated by the disappearance of this particle through the inverse reaction, which would occur with the same frequency in equilibrium.

If these three conditions are present, then it is possible to produce a non-vanishing global baryon number. The question is now to consider specific models in which we can quantify the amplitude of this effect, and compare it with observations.

9.3.1.1 Observations

Denoting by $N_B = n_b - \bar{n}_b$ the asymmetry between baryons, n_b , and antibaryons, \bar{n}_b , and n_γ the number density of photons, then observations concerning baryogenesis usually imply the ratio [see (4.119)]

$$\eta \equiv \frac{N_B}{n_\gamma} \simeq 5 - 6 \times 10^{-10}, \quad (9.30)$$

which is a small, but non-vanishing number that the theories should explain.

9.3.1.2 Quantum gravity

Regardless of the precise theory, one can assume that quantum gravity effects, during the very early moments of the Universe, at a temperature of the order of the Planck temperature, are capable of violating any quantum number conservation, and as a consequence, B , as well as \hat{CP} . Since the notion of thermal equilibrium is probably not defined for these theories, the Sakharov set of conditions can be hypothetically assumed to be satisfied and baryogenesis could have happened before anything else.

Unfortunately, this vision is not very precise since, in the absence of any convincing theory, it is impossible to make any calculation at all. So, in some sense, it essentially reduces to imposing an initial condition, in a not so much more elaborate way (but a better wording!). Such baryogenesis at the Planck scale also has a major inconvenience. As explained in Chapter 8, the Universe has most probably undergone an inflationary phase during which any pre-existing asymmetry would have been washed out. This is why in the present state, credible baryogenesis mechanisms are based on energy scales either of the order of the grand unification scale, or the electroweak scale (or perhaps a mixture of both).

9.3.1.3 Grand unification

Almost all grand unification theories couple quarks and leptons in such a way as to effectively violate the conservation of baryon number, making them serious candidates for baryogenesis. Furthermore, while even the standard electroweak model predicts a \hat{CP} violation (which has been observed), there is no reason why a realistic grand

unification model, which should also describe these effects, not also violate \hat{CP} , and presumably at a higher level (this has been the case of all the theories that have been studied in detail so far). Finally, many particles in these theories are unstable, starting with the gauge bosons X and Y . Their decay, at the time where the temperature becomes small enough to stop their production in sufficient number, breaks the thermodynamical equilibrium. Moreover, these particles are responsible for the non-conservation of B . In other words, Sakharov's conditions can here again be satisfied, and precise calculations seem to indicate that the produced baryon number could be compatible with observations [9].

In practice, we assume that as long as the temperature is greater than the energy scale E_{GUT} , the Universe is in thermal equilibrium. Then, comes the time at which the production of X and Y becomes too small to compensate their decay, i.e. when the expansion rate H of the Universe becomes comparable to the lifetime of these particles, (see Chapter 4). The $\hat{CP}\hat{T}$ theorem then implies that during this period, as many X as \bar{X} must be produced since their total decay width must be precisely equal. Having said that, the partial decay widths, i.e. the decay probabilities in each possible channel, can be different. So, for instance, the channels

$$(X \rightarrow d \ell ; X \rightarrow \bar{q} \bar{u}),$$

where ℓ (q) represents an arbitrary element of the lepton (quark) doublet of $SU(2)_L$, are not necessarily equal to the channels

$$(\bar{X} \rightarrow \bar{d} \bar{\ell} ; \bar{X} \rightarrow q u),$$

although, in general, it is quite a small effect (second order in perturbations, at one loop). In this case, the production of quarks and antiquarks will be slightly different, leaving a small relic density.

There are strong reasons to think that these mechanisms, although simple to apply, were not responsible for primordial baryogenesis. Just as for the Planck scale, the inflationary phase that follows inevitably tends to dilute all the asymmetry. This inflationary phase is followed by a reheating during which numerous particles are produced, and that rapidly enter thermal equilibrium. For reasons explained in Chapter 11 (over-production of monopoles), the temperature reached must absolutely be lower than the phase transition temperature of grand unification, E_{GUT} . As a result, this mechanism cannot have been efficient after inflation. So it cannot produce the baryon asymmetry that is observed today.

9.3.2 Electroweak anomalies

At the classical and perturbative level, the baryon and lepton numbers are conserved separately by the electromagnetic and weak interactions. However, at the quantum level, the theory possesses what are called anomalies, so that the currents are not conserved. This is expressed by the relation

$$\partial_\mu J_B^\mu = \partial_\mu J_L^\mu = \frac{N_F}{8\pi^2} \text{Tr} B_{\mu\nu} \tilde{B}^{\mu\nu}, \quad (9.31)$$

with a number $N_F = 3$ of families. This relation replaces the usual conservation, in the case of baryonic and leptonic currents. We notice that the $B - L$ current is conserved: this current also appears in unifications of the SU(5) or SO(10) type.

In equation (9.31), we formed matrix fields $B_{\mu\nu}$ and their dual $\tilde{B}^{\mu\nu}$ on the basis of the vectors of $SU(2)_L$ [see (2.151)] using the definitions [where $B_{i\mu\nu}$ is the Faraday tensor of the group $SU(2)_L$, identifiable in the Lagrangian (2.154)]

$$B_{\mu\nu} \equiv \frac{1}{2} \sum_i B_{i\mu\nu} \sigma^i, \quad \text{and} \quad \tilde{B}^{\mu\nu} \equiv \frac{1}{2} \varepsilon^{\mu\nu\alpha\beta} B_{\alpha\beta}, \quad (9.32)$$

which generalizes easily to the case of a group with arbitrary generators T^i by performing the substitution $\frac{1}{2}\sigma^i \mapsto T^i$.

9.3.2.1 Chern–Simons number

Although apparently complicated, the expression on the right-hand side of (9.31) allows for a considerable simplification as soon as we notice the identity

$$\text{Tr } B_{\mu\nu} \tilde{B}^{\mu\nu} = \partial_\mu \left\{ \varepsilon^{\mu\nu\alpha\beta} \text{Tr} \left[\left(B_{\nu\alpha} + \frac{2}{3} B_\nu B_\alpha \right) B_\beta \right] \right\}, \quad (9.33)$$

where $B_\mu = \frac{1}{2} B_{i\mu} \sigma^i$.

Denoting by $B = \int d^3x \mathcal{J}_B^\mu$ the baryon number associated with the current \mathcal{J}_B^μ and using the fact that all configurations of physical interest vanish sufficiently fast at infinity, we find that between the times t_1 and t_2 , the variation of the baryon number ΔB is given by

$$\Delta B = N_F \Delta N_{\text{CS}} = N_F [N_{\text{CS}}(t_2) - N_{\text{CS}}(t_1)]. \quad (9.34)$$

The Chern–Simons numbers, defined by

$$N_{\text{CS}} = \frac{1}{4\pi^2} \int d^3x \varepsilon^{ijk} \text{Tr} \left[B_i \left(\partial_j B_k + \frac{4}{3} B_j B_k \right) \right], \quad (9.35)$$

end up being integer numbers as long as the field configuration is pure gauge, i.e. if we can write $B = -iU^{-1}\nabla U$, where U is a gauge transformation.

9.3.2.2 Instanton and sphaleron

At the perturbative level, two pure gauge field configurations are in principle indistinguishable. Both of them can be used as a vacuum state for the theory, up to the fact that they do not have the same Chern – Simons number. There can exist a Euclidean solution, called an *instanton*, which interpolates between the two configurations. Knowing the structure of an instanton, we can compute the probability to tunnel from one vacuum state to another, which gives the probability to produce a baryon number spontaneously. The typical expected magnitude of this effect is of the order of $e^{-1/g_s^2} \sim 10^{-91}$ at low energy ($T \ll M_W$) and is thus completely negligible in any non-cosmological situation.

The global structure of the configuration space has then an infinity of minima, one for each value of $N_{\text{CS}} \in \mathbb{Z}$ corresponding to the (pure gauge) vacuum configurations

related with instantons. The height of the potential barrier that separates these states can be evaluated by studying the time-independent configuration that corresponds to the energy maximum. Such a solution was found by Klinkhamer and Manton [12] and is called a *sphaleron*. This solution, which is necessarily unstable, has an energy of the order of

$$E_{\text{sphaleron}} \propto \frac{M_W}{g_2^2},$$

and it was found that, via the sphaleron mode, the transition width per unit volume for a unit Chern–Simons variation [transition of $\Delta B = 3$ in the standard model, so following (9.34)], is of the order of

$$\frac{\Gamma_{\text{sphaleron}}}{V} \simeq T^4 \exp\left(\frac{E_{\text{sphaleron}}}{T}\right),$$

when the Universe is at a temperature T . When the temperature reaches that of the electroweak transition, the mass of the intermediate boson decreases, so that this last result can a priori no longer apply. Actually, it turns out that it leads to a good order of magnitude since the best estimate that we find in this case is $\Gamma/V \simeq T^4/g_2^8$.

9.3.3 Electroweak baryogenesis

9.3.3.1 $B - L$ conservation

It is important to realize that even though the standard model of the electroweak interactions has in itself a mechanism capable of producing a baryon number dynamically, it does so while conserving the combination $B - L$. This implies that the sphaleron can also act in the opposite way: if an arbitrary $B - L$ conserving phenomenon produces a baryon number prior to the electroweak transition, then, due to the sphaleron, this transition will automatically wash out such a primordial contribution. This is a very constraining restriction for models of baryogenesis using the grand unification transition.

In the present form of the standard model, its contribution to baryogenesis is largely insufficient. Plausible extensions are, however, possible and should not be disregarded. Therefore, these effects must be taken into account since they lead to a possible electroweak scale contribution to baryogenesis.

9.3.3.2 \hat{CP} violation

As shown in equations (9.16) and (9.17), the \hat{CP} symmetry is not respected by the electroweak interactions due to the non-trivial phase δ_{CKM} of the CKM matrix (see Chapter 2). Unfortunately, this phase is very small: its value can be estimated by computing explicitly the transition probabilities between states with different \hat{CP} values. Using the known experimental values (Ref. [11] of Chapter 2) and the standard model, we find

$$\delta_{\text{CKM}} \lesssim 10^{-25},$$

knowing that this value is a pre-factor for any estimate of the production of B based on the sphaleron. In other words, while a production of $\Delta B = N_f \Delta N_{\text{CS}}$ is already weak, since $\Delta N_{\text{CS}} = 1$ is very favoured, this number is considerably reduced by the phase factor δ_{CKM} .

9.3.3.3 Electroweak transition

The last blow for the mechanism of electroweak baryogenesis comes from the transition itself. In order for the production of B to simply be possible, the transition should be out of equilibrium. To compensate the weak \hat{CP} violation, the transition should be as far from equilibrium as possible. For this, one needs a first-order transition, i.e. a cubic term in the effective potential of the Higgs field at finite temperature. But it turns out that when computing this cubic term as a function of the masses of the Z , the W and the Higgs, we find quite a small coefficient, of the order of 10^{-2} at best, when considering the upper limit of the Higgs mass.

Since the transition is weakly first order, it happens by nucleation. Bubbles of true vacuum are expanding in the space otherwise still in the false vacuum. The Higgs fields vary rapidly on the surface of these bubbles, creating the required conditions for baryogenesis. It is thus necessary, in order to produce a significant number of baryons, that the production process stops rapidly in the true vacuum after the bubble passage, so as to freeze the number produced. If these processes were still happening in the true vacuum phase, which is in thermodynamical equilibrium, then antibaryons would be rapidly produced to compensate the initial baryons and the initial asymmetry would be lost.

In order for the $B - L$ violation to be weak in the true vacuum after the transition, the energy of the sphaleron should be very large compared to the temperature right after the bubble surface, i.e. the intermediate bosons should have significant masses, and the expectation value of the Higgs field should grow very rapidly. This depends on the coefficients obtained in the effective potential and on the Higgs mass at zero temperature. Given the constraints on the latter ($M_H \gtrsim 115$ GeV), we find that the sphaleron transition, which produces the required baryon number across the bubble surface, reduces it drastically in the subsequent true vacuum phase. The electroweak theory is therefore not sufficient for baryogenesis. This is even more true as indicated from recent calculations that seem to indicate that when higher order perturbative corrections are taken into account, the transition is not even weakly first order, but merely a smooth cross-over, i.e. no transition as such! This is a real ‘coup de grâce’ for electroweak baryogenesis, which implies that the standard model must be extended somehow.

9.3.3.4 Modifications

Some models [9] have been proposed according to which baryogenesis can be realized at the electroweak transition. These models all share the common property of modifying the low-energy physics by high-energy-inducing effects. Supersymmetry, for instance, introduces new particles, which modifies the coefficients of the Higgs effective potential, and as a result can make the transition more strongly first order. These new particles imply new interactions, and as a consequence the possibility of more strongly violating the \hat{CP} symmetry. For now, the space of parameters that allows for a sufficient baryogenesis is quite constrained, and does not seem very natural, so that these hypotheses are not retained as being the most favourable ones.

Another possibility to study is that of a non-thermal mechanism such as preheating after the inflationary phase (Chapter 8). If effects of the sphaleron type exist during

the preheating phase, then the production could be considerably amplified. In other words, and even if this kind of scenario is not favoured by current research, it is still possible that the electroweak transition could have contributed to baryogenesis.

We finally note that the possibility of an non-thermal mechanism also applies to the models described above.

9.3.4 Leptogenesis

9.3.4.1 Back to the ‘see-saw’ mechanism

As explained above, one of the direct, and unavoidable, consequences of grand unified theories is the existence of a neutrino mass. This mass is generated by the ‘see-saw’ mechanism which makes it possible to obtain very different masses between the left- and right-handed components of the neutrino. Where can a mass term such as that of equation (9.27) arise from? As for all mass terms, its most likely origin lies in a Yukawa coupling with the Higgs mass. We can thus write

$$\mathcal{M} = |\phi| \lambda,$$

where λ is a coupling constant matrix and the constant part \mathcal{M} of the matrix comes, in this case, from the non-vanishing expectation value of the Higgs field.

Since the sterile neutrino does not have any quantum number, the reinterpretation of (9.27) as an interaction with exchange of a Higgs boson indicates that this sterile neutrino can decay into a Higgs boson together with an antilepton or into an antiHiggs together with a lepton. The lepton number is thus no longer conserved. Moreover, together with the massive neutrinos come new phases in the mass matrix and in this way, the $\hat{\mathcal{CP}}$ violation can be much more significant than in the quark sector. As a result, and identically to models based on grand unification, the probabilities of decay into neutrinos or antineutrinos can be different, giving rise to leptogenesis.

The scenario will thus be the following [9]. Because of either thermal or non-thermal reasons, at a given time in its history the Universe contains a large number of right-handed neutrinos. The thermal reason can be that the temperature of the Universe simply becomes lower than the mass of the right-handed neutrinos. When a non-thermal mechanism is at work, there can be a parametric resonance production of a large number of right-handed neutrinos. In both cases, right-handed neutrinos being unstable, they start decaying, producing a global lepton number due to the different cross-sections into the neutrino or antineutrino modes. This produces a global $B - L$ since there is no production of baryons associated with leptogenesis. The amplitude is computed as a function of the branching ratios of the sterile neutrino decay. This initial $B - L$, actually due to this initial lepton number, is then converted into a baryon number through the process related to the sphaleron. One then finds [11]

$$B = \frac{4(N_F + 2)}{22N_F + 13} (B - L)_{\text{ini}},$$

where $(B - L)_{\text{ini}}$ is the lepton number initially produced during the neutrino decays. For ‘reasonable’ values of the microscopic parameters, actually undetermined, this leads to a baryogenesis in agreement with the observational constraint (9.30).

It is important to notice that this mechanism is not only based on the existence of a neutrino mass, which is now almost established. It is also completely natural in the context of supersymmetric grand unified theories with hybrid inflation, models that turn out to be very favoured for completely independent reasons.

9.3.5 Affleck–Dine mechanism

Another mechanism that works well relies on the coherent oscillations of a scalar field, proposed by Affleck and Dine in 1985 [13], and requires that the field has a baryonic charge.

9.3.5.1 A baryonic scalar field

The Affleck–Dine mechanism for baryon-number production relies on the possibility that a complex scalar field ϕ , such as the one described by the Lagrangian (2.103), carries a baryon number, i.e. that the $U(1)$ symmetry associated with it is precisely the one whose hypercharge is B . In this case, the current (2.111) belongs to the total baryonic current. Moreover, the Lagrangian (2.103) is also invariant under the transformation $\phi \leftrightarrow \phi^*$, which can be interpreted as a \hat{CP} symmetry [cf. the transformation laws (2.170) and (2.175) for the scalar field].

This kind of field is expected in the supersymmetric extensions of particle theories. In this framework, discussed in full detail in Chapter 10, to each fermion one associates a bosonic partner with the same quantum numbers, and demands a new symmetry, called supersymmetry, to hold between these fermions and bosons. Similarly, in fact for the symmetry to be possible, one also associates a fermion to each boson. Since the baryon and lepton numbers of the quarks and leptons are determined by their quantum numbers, then those of the supersymmetric partners will also be. There must therefore be some scalar fields carrying leptonic and baryonic charges.

9.3.5.2 Coherent oscillations

Scalar fields, as shown in Section 8.6 in Chapter 8 on the example of preheating, can oscillate coherently. In a supersymmetric extension such as discussed above, some scalar fields do carry leptonic and baryonic charges, so that if they were made to oscillate coherently, they could produce large baryon and lepton numbers.

During a phase of coherent oscillations, a free massive scalar field behaves, on average, as a dust fluid. In particular, the test field evolves as

$$a^{-3/2}\phi \propto \sqrt{t}Z_{(3\alpha-1)/2}(mt) \quad (9.36)$$

in a Universe where the scale factor evolves as $a \propto t^\alpha$; in (9.36), $Z_{(3\alpha-1)/2}$ represents a generic Bessel function of index $(3\alpha - 1)/2$.

At very short times, one can expand the Bessel function, leading to a constant mode ϕ_0 that dominates the evolution of the scalar field. At late times, $mt \gg 1$, the field behaves as $\phi \propto a^{-3/2} \sin(mt + \beta)$, where β is an arbitrary phase. This is the phase of coherent oscillation that is of interest for particle production: for this, we need to add another ingredient in the model, namely we introduce interactions.

9.3.5.3 Particle production

Let us now consider self-interacting terms for ϕ which neither conserve the baryon number nor respect $\hat{C}\hat{P}$. Consider a potential of the form

$$V(\phi) = \lambda|\phi|^4 + (\epsilon\phi^3\phi^* + \delta\phi^4 + \text{h.c.}), \quad (9.37)$$

in which the coupling constants, ϵ and δ , are small and complex so as to violate $\hat{C}\hat{P}$. Only the first term of (9.37) conserves the baryon number.

We now assume that at the initial time, the scalar field is real, i.e. $\phi_0 \in \mathbb{R}$. Neglecting all the contributions from the imaginary part compared to the real part, $\Im(\phi) \ll \Re(\phi) \simeq \phi_0$, the equation of motion of the imaginary part is

$$\frac{d^2}{dt^2}\Im(\phi) + 3H\frac{d}{dt}\Im(\phi) + m^2\Im(\phi) \simeq \Im(\epsilon + 2\delta)\Re(\phi). \quad (9.38)$$

The solution of (9.38) for $mt \ll 1$ is obtained by writing in this regime $\Re(\phi) = \phi_0$, which fixes the amplitude of the imaginary part. When $mt \gg 1$ on the other hand, the source term becomes rapidly negligible and we find the same solution as for the real part,

$$\Im(\phi) \propto \Im(\epsilon + 2\delta) a^{-3/2} \sin(mt + \tilde{\beta}),$$

where $\tilde{\beta}$ is still an arbitrary constant that we compute numerically, just as for the coefficients of proportionality.

With the solution for the real and imaginary parts of the scalar field, it is sufficient to compute the time component of the current to obtain the number density of baryons produced through this mechanism. We find

$$n_B \propto \Im(\epsilon + 2\delta) a^{-3} \sin(\beta - \tilde{\beta}),$$

which shows explicitly that a baryon number can be produced provided the two following conditions are satisfied:

(1) The Lagrangian must contain a term that violates the conservation of the baryon number, i.e. it is required that $\epsilon \neq 0$ or $\delta \neq 0$.

(2) The coupling constants must have non-vanishing imaginary parts.

It is interesting to note at this point that the rigorous implementation of models of the Affleck–Dine type in supersymmetric extensions of the standard model or in supersymmetric theories of grand unification is possible and even natural, and leads to values of η compatible with (9.30).

There is no generally accepted mechanism for baryogenesis, and much work still needs be done on this topic [14]. As bizarre as it may sound, we find it rather encouraging, however, to realize that no standard model physics model has ever been able to produce baryons at a sufficient level: since the asymmetry between baryons and antibaryons is an observed fact, some sort of new physics is at work there, and not necessarily at unreachable energy scales. Some new models to produce this asymmetry will thus hopefully be testable not only with the forthcoming better observations, but also through direct production, in accelerators, i.e. experiments, of the new particles they will unavoidably predict.

References

- [1] S. TOMONAGA, ‘On a relativistically invariant formulation of the quantum theory of wave fields’, *Prog. Theor. Phys.* **1**, 27, 1946; J. SCHWINGER, ‘Quantum Electrodynamics. I. A Covariant Formulation’, *Phys. Rev.* **74**, 1439, 1948.
- [2] R.D. MCKEOWN and P. VOGEL, ‘Neutrinos masses and oscillations, triumphs and challenges’, *Phys. Rep.* **394**, 315, 2004; data from KamLAND of Fig. 9.1 are analysed by the collaboration, ‘Measurement of Neutrino Oscillation with KamLAND: Evidence of Spectral Distortion’, *Phys. Rev. Lett.* **94**, 081801, 2005.
- [3] The official site at <http://www-sk.icrr.u-tokyo.ac.jp/sk/index-e.html> provides more technical details and results.
- [4] Their official web site at <http://kamland.lbl.gov/> provides the latest results.
- [5] H. GEORGI and S.L. GLASHOW, ‘Unity of All Elementary-Particle Forces’, *Phys. Rev. Lett.* **32**, 438, 1974.
- [6] J.C. PATI and A. SALAM, ‘Unified lepton-hadron symmetry and a gauge theory of basic interactions’, *Phys. Rev. D* **8**, 1240, 1973.
- [7] P. LANGACKER, ‘Grand Unified Theories and Proton decay’, *Phys. Rep.* **72**, 185, 1981.
- [8] *The BaBar Physics Book*, rapport SLAC-R-504 (2001),
<http://www.slac.stanford.edu/pubs/slacreports/slac-r-504.html>.
- [9] M. DINE and A. KUSENKO, ‘The origin of the matter-antimatter asymmetry’, *Rev. Mod. Phys.* **76**, 1, 2004. A summary on the main notions is also found in the review article of A. RIOTTO and M. TRODDEN, ‘Recent Progress in Baryogenesis’, *Ann. Rev. Nucl. Part. Sci.* **49**, 35, 1999.
- [10] A. D. SAKHAROV, ‘Violation of CP Invariance, C Asymmetry, and Baryon Asymmetry of the Universe’, *ZhETF Pis ma Redaktsiiu* **5**, 32 1967.
- [11] J.A. HARVEY and M.S. TURNER, ‘Cosmological baryon and lepton number in the presence of electroweak fermion-number violation’, *Phys. Rev. D* **42**, 3344, 1990.
- [12] F. KLINKHAMER and N. MANTON, ‘A saddle-point solution in the Weinberg-Salam theory’, *Phys. Rev. D* **30**, 2212, 1984.
- [13] I. AFFLECK and M. DINE, ‘A new mechanism for baryogenesis’, *Nucl. Phys. B* **249**, 361, 1985.
- [14] W. BUCHMÜLLER, ‘Baryogenesis – 40 Years Later’ Proceedings PASCOS - 07, [arXiv:0710.5857](https://arxiv.org/abs/0710.5857) 2007.

10

Extensions of the theoretical framework

The different theoretical frameworks we have elaborated on so far need to be extended since they can be recognized as being incomplete in two ways. First, gravity is never included in particle-physics theories, which is justified inasmuch as it does not have significant effects in the quantum interactions of these particles, at least as long as the energies involved are small compared to the Planck scale.

For this reason we shall begin by studying the modifications that arise from the couplings between classical fields, for instance, scalar fields, and gravity: these are the scalar-tensor theories (Section 10.1). We then consider particle physics in an arbitrary curved space-time (Section 10.2).

Second, the particle-physics framework itself does not seem to be complete. For instance, as clearly shown in Fig. 9.2, the unification of the three coupling constants at high energies, whilst seemingly plausible, is not definitive from these measurements. Supersymmetry, which will be discussed in Section 10.3, allows for a much better unification (see Fig. 10.5). The natural extension of this supersymmetry arises when we impose it to be a local symmetry; in this case we note that space-time transformations become gauge quantities. In other words, gravity appears spontaneously in this framework, which is conveniently referred to as supergravity (Section 10.3.6). Finally, supersymmetry is a prediction of superstring theory, which is a supposedly fundamental theory on which a substantial part of research on fundamental physics is focused (Chapter 13).

10.1 Scalar-tensor theory of gravity

Scalar-tensor theories are the most natural extensions of the theory of general relativity. In these theories, gravity is mediated by a massless spin-2 field, the graviton, $g_{\mu\nu}$, just as in general relativity, and by one or several spin-0 scalar fields, φ . These theories were initially considered by Jordan, Fierz, Brans and Dicke [1].

New scalar degrees of freedom generically appear in theories with extra dimensions (Chapter 13), in particular in string theory. In this latter case, the so-called dilaton is present in the supermultiplet of the 10-dimensional graviton and other scalar fields, called *moduli* (see Section 13.2.5.3), appear as a Kaluza–Klein-like compactification to a four-dimensional space-time is performed.

From a field-theory point of view, scalar-tensor theories merely describe a scalar field universally coupled to matter so that it respects most of the symmetries of general relativity: conservation laws, the constancy of the non-gravitational constants (see

Section 12.3 of Chapter 12) and local Lorentz invariance. Moreover, the universality of free-fall is also ensured as long as the gravitational binding energy is negligible. Thus, it is a well defined and theoretically well-motivated extension of general relativity, offering an interesting phenomenology for cosmology. Notice that more general scalar-tensor theories, such as those that arise in the context of Kaluza–Klein theories or from the dilaton of string theory (see Chapter 13), violate the constancy of the non-gravitational constants because these fields are not universally coupled to matter.

10.1.1 Formulation

To describe the properties of these models, let us consider the case of a theory with a single scalar field (for a more general description, see Ref. [2]).

10.1.1.1 Jordan frame

In the so-called *Jordan frame* (also often referred to as ‘string’ frame), one simply assumes that there exists a scalar field that is coupled to the Ricci scalar. The action of the theory then takes the general form

$$S = \frac{1}{16\pi G_*} \int d^4x \sqrt{-g} [F(\varphi)R - g^{\mu\nu}Z(\varphi)\varphi_{,\mu}\varphi_{,\nu} - 2U(\varphi)] + S_m[g_{\mu\nu}; \text{matter}]. \quad (10.1)$$

where G_* is the ‘bare’ gravitational constant (which, as we shall see, differs from the measured one, G_N) from which we define $\kappa_* = 8\pi G_*$ and S_m is the action for the matter fields. The latter are minimally coupled, i.e. with no specific additional coupling to the metric $g_{\mu\nu}$, so that it is only a functional of the matter fields and $g_{\mu\nu}$ and does not involve the scalar field φ . Note that, because of the overall factor G_* , φ is dimensionless. The theory seems to depend on three arbitrary functions, two of which, F and Z , are dimensionless, and U is the scalar field potential. The function F needs to be positive-definite for the graviton to carry positive energy. One can always absorb one of the two functions F or Z in a redefinition of the scalar field so that the theory effectively has only two arbitrary functions.

Two parameterizations are often used,

- the Brans–Dicke parameterization where we set

$$F(\varphi) = \varphi, \quad Z(\varphi) = \frac{\omega(\varphi)}{\varphi}, \quad (10.2)$$

- the simplest parameterization, which can sometimes be pathological, for which

$$F(\varphi) \quad \text{arbitrary}, \quad Z(\varphi) = 1. \quad (10.3)$$

In the Jordan frame, the matter fields are universally coupled to the metric tensor, $g_{\mu\nu}$, so that our lab clocks and rulers, which are composed of ordinary matter, are not affected by the local value of the scalar field. Therefore in this frame, any experimental measurement will have its usual interpretation. This will, for instance, be the case for the Hubble constant, the redshifts, etc. Moreover, all physical processes, such as nuclear processes, take their usual form in this frame.

In the most general case, varying the action (10.1) with respect to the metric gives the (modified) Einstein equations,

$$\begin{aligned} F(\varphi)G_{\mu\nu} &= 8\pi G_* T_{\mu\nu} + Z(\varphi) \left[\partial_\mu \varphi \partial_\nu \varphi - \frac{1}{2} g_{\mu\nu} (\partial_\alpha \varphi)^2 \right] \\ &\quad + \nabla_\mu \partial_\nu F(\varphi) - g_{\mu\nu} \square F(\varphi) - g_{\mu\nu} U(\varphi), \end{aligned} \quad (10.4)$$

where ∇_μ is the covariant derivative associated with $g_{\mu\nu}$ and we use the notation $F_{,\varphi} = dF/d\varphi$. The energy-momentum tensor of the matter is defined by (1.84), as usual.

Varying the action (10.1) with respect to φ leads to the Klein–Gordon equation governing the evolution of the scalar field,

$$Z(\varphi) \square \varphi = U_{,\varphi} - \frac{1}{2} F_{,\varphi} R - \frac{1}{2} Z_{,\varphi} (\partial_\alpha \varphi)^2. \quad (10.5)$$

Finally, variations with respect to the matter fields reduce to the standard conservation equations, since there is no coupling between these fields and φ , namely

$$\nabla_\mu T^{\mu\nu} = 0. \quad (10.6)$$

The Klein–Gordon equation (10.5) can be rewritten using the trace of (10.4) to determine the scalar curvature, R . We then obtain an evolution equation for the scalar field in the Brans–Dicke form,

$$2\varpi \square \varphi = 8\pi G_* F_{,\varphi} T - \varpi_{,\varphi} (\partial_\alpha \varphi)^2 - 4F_{,\varphi} U + 2U_{,\varphi} F, \quad (10.7)$$

with $T \equiv T_{\mu\nu} g^{\mu\nu}$ and $2\varpi \equiv 2ZF + 3F_{,\varphi}^2$. This function reduces to $2\omega(\varphi) + 3$ in the Brans–Dicke representation.

10.1.1.2 Einstein frame

From a theoretical point of view, it is interesting to study these theories in the Einstein frame defined by diagonalizing the kinetic terms for the graviton and scalar field. This is achieved thanks to a ‘conformal’ transformation of the metric,

$$g_{\mu\nu}^* = F(\varphi) g_{\mu\nu}, \quad (10.8)$$

and a redefinition of the scalar field according to

$$\left(\frac{d\varphi_*}{d\varphi} \right)^2 = \frac{3}{4} \left[\frac{d \ln F(\varphi)}{d\varphi} \right]^2 + \frac{Z(\varphi)}{2F(\varphi)}, \quad (10.9)$$

$$A(\varphi_*) = F^{-1/2}(\varphi), \quad (10.10)$$

$$2V(\varphi_*) = U(\varphi)F^{-2}(\varphi). \quad (10.11)$$

Using these new variables and functions, the action (10.1) takes the form

$$S = \frac{1}{16\pi G_*} \int d^4x \sqrt{-g_*} [R_* - 2g_*^{\mu\nu} \partial_\mu \varphi_* \partial_\nu \varphi_* - 4V(\varphi_*)] + S_m[A^2(\varphi_*)g_{\mu\nu}^*; \text{matter}]. \quad (10.12)$$

where g_* is the determinant of the metric $g_{\mu\nu}^*$ and R_* its Ricci scalar. More generally, any quantity with a star (*) will be assumed to be defined in this frame. As it is

written, the first term is exactly the gravitational action of general relativity plus a minimally coupled scalar field φ_* but now the matter fields are explicitly coupled to this scalar field through the coupling function $A^2(\varphi_*)$.

The Einstein, Klein–Gordon and matter-conservation equations, obtained by varying the action (10.12) with respect to the various fields, then take the form

$$G_{\mu\nu}^* = 8\pi G_* T_{\mu\nu}^* + 2\partial_\mu \varphi_* \partial_\nu \varphi_* - g_{\mu\nu}^* (\partial_\alpha \varphi_*)^2 - 2g_{\mu\nu}^* V, \quad (10.13)$$

$$\square_* \varphi_* = V_{,\varphi_*} - 4\pi G_* \alpha(\varphi_*) T_{\mu\nu}^* g_*^{\mu\nu}, \quad (10.14)$$

$$\nabla_\mu T_*^{\mu\nu} = \alpha(\varphi_*) T_{\sigma\rho}^* g_*^{\sigma\rho} \partial^\nu \varphi_*, \quad (10.15)$$

where the function α , which is the coupling strength of the scalar field to matter sources, is defined by

$$\alpha(\varphi_*) \equiv \frac{d \ln A}{d \varphi_*}, \quad (10.16)$$

whose derivative

$$\beta(\varphi_*) \equiv \frac{d\alpha}{d\varphi_*} \quad (10.17)$$

will also be needed later on.

The stress-energy tensor $T_{\mu\nu}^*$ is now defined from the variation of the matter action with respect to $g_{\mu\nu}^*$ as

$$T_*^{\mu\nu} \equiv \frac{2}{\sqrt{-g_*}} \frac{\delta S_m}{\delta g_{\mu\nu}^*},$$

and one can check that it is related to the energy-momentum tensor defined in the Jordan frame by $T_{\mu\nu}^* = A^2 T_{\mu\nu}$.

The advantage of the Einstein frame lies in the fact that the transformation (10.8) diagonalizes the kinetic terms of the graviton and scalar field. So, linearizing around a Minkowski space-time, the spin 0 and 2 degrees of freedom, i.e. scalar and tensor, are the perturbations of φ_* and $g_{\mu\nu}^*$ (whose kinetic term is the standard Einstein–Hilbert one), respectively. The field equations (10.13)–(10.15) are written such that all the second derivatives of the fields are gathered on the left-hand side, which is not the case for (10.4) and (10.5). The Cauchy problem is therefore well posed in the Einstein frame. So, the theory is mathematically well defined only if we can write the action in the Einstein frame. In particular, the kinetic term of the scalar field φ_* should have a negative sign in order for it to carry a positive energy. Moreover, if the transformations (10.9)–(10.11) are singular for some values of φ , the consistency of the theory should be studied in the Einstein frame. Various situations can arise: (1) some apparent singularities may only be artifacts of the parameterization (10.1) and are consequently not physical, (2) quantities in the Jordan frame may look to be well defined, while there is a singularity in Einstein frame (a typical case is when F vanishes) and the model must then be considered as pathological.

10.1.1.3 Gravitational constants

The action (10.1) in the Einstein frame is used to define an effective gravitational constant as

$$G_{\text{eff}} \equiv \frac{G_*}{F(\varphi)}. \quad (10.18)$$

However, this quantity does not represent the value of the gravitational constant that would be measured in a Cavendish-type experiment. In such an experiment, the amplitude of the Newton force, F , between two masses m_1 and m_2 is measured. The gravitational constant is then defined by $G_{\text{cav}} = Fr^2/m_1m_2$. If the scalar field is light, it contributes to the force F at large distances, since the mass is coupled to both the graviton and the spin-0 degree of freedom. It can then be shown that

$$G_{\text{cav}} = G_* A^2 (1 + \alpha^2) = \frac{G_*}{F} \left(\frac{2ZF + 4F_{,\varphi}^2}{2ZF + 3F_{,\varphi}^2} \right). \quad (10.19)$$

Indeed, the Newton constant is simply G_{cav} today, i.e. $G_N = G_{\text{cav}}(\varphi_0)$. In this expression, the first term ($G_* A^2$) corresponds to the exchange of a graviton between the two bodies while the second term ($G_* A^2 \alpha^2$) corresponds to the exchange of a scalar between them. In the Brans–Dicke representation one has, $G_{\text{cav}} = G_* \varphi^{-1} (2\omega + 4)/(2\omega + 3)$.

Various remarks are in order:

- Expression (10.19) implies that the gravitational constant measured in the lab, or in the Solar System, is a dynamical quantity. It can thus vary in space and time. Current constraints [3] imply that

$$\left| \frac{\dot{G}_N}{G_N} \right|_0 < 6 \times 10^{-12} \text{ year}^{-1}. \quad (10.20)$$

- In the Jordan frame, the gravitational constant is variable, while the particle masses are fixed. In the Einstein frame, the situation is inverted. Nevertheless, as emphasized above, one cannot distinguish between the two frames inasmuch as only the dimensionless quantity $\alpha_G \equiv G_N m^2/\hbar c$ is accessible experimentally.
- General relativity is now represented by a point in the space of scalar-tensor theories, namely $(\alpha, \beta) = (0, 0)$.
- If the dilaton has a mass m_* in the Einstein frame, then the scalar force has a range of the order of $\lambda = m_*^{-1}$. For distances smaller than m_*^{-1} the scalar contribution is significant and the gravitational constant is given by (10.19) whereas for larger distances, the contribution from the scalar field is negligible and the theory reduces to general relativity. At the Newtonian level, the gravitational potential takes the form

$$V(r) = -G_N \left(1 + \alpha_{12} e^{-r/\lambda} \right) \frac{m_1 m_2}{r}, \quad (10.21)$$

describing a Yukawa deviation to the standard Newtonian potential. The amplitude of the deviation is given by $\alpha_{12} = f_1 f_2$, where f_i is defined by

$$f_i \equiv \frac{\partial \ln m_i(\varphi_*)}{\partial \varphi_*}, \quad (10.22)$$

and is universal; the functions $m_i(\varphi_*)$ do not depend on the chemical compositions of the objects. The parameters (α_{12}, λ) are constrained experimentally in the Solar System (see Fig. 1.11).

10.1.2 Local constraints

The predictions of general relativity in the weak-field limit are confirmed by Solar-System experiments to a high accuracy. Scalar-tensor theories must satisfy these constraints and thus be close to general relativity today. Figure 1.9 summarizes the constraints on the post-Newtonian parameters γ^{PPN} and β^{PPN} defined by (1.140).

For scalar-tensor theories, these two parameters are given explicitly [2, 4, 5] by

$$\gamma^{\text{PPN}} - 1 = -2 \frac{\alpha^2}{1 + \alpha^2} = -\frac{F_{,\varphi}^2}{ZF + 2F_{,\varphi}^2}, \quad (10.23)$$

$$\beta^{\text{PPN}} - 1 = \frac{1}{2} \frac{\alpha^2 \beta}{(1 + \alpha^2)^2} = \frac{1}{4} \frac{FF_{,\varphi}}{2ZF + 3F_{,\varphi}^2} \frac{d\gamma^{\text{PPN}}}{d\varphi}, \quad (10.24)$$

respectively, in the Einstein and Jordan frames. The observational constraints of Fig. 1.9 can be translated in the plane (α_0, β_0) to obtain the constraints summarized in Fig. 10.1. In the Solar System, any scalar-tensor theory is completely defined by these two numbers only [2, 4]. The Solar-System constraints imply that

$$\alpha_0^2 < 4 \times 10^{-5}, \quad (10.25)$$

while the observation of binary pulsars provide [6] that

$$\beta_0 > -4.5. \quad (10.26)$$

The constraint (10.25) implies that today G_{eff} and G_{N} , or equivalently G_{cav} , do not differ by more than $10^{-3}\%$. However, notice that on cosmological scales, this difference can be much more important. Equation (10.20) also implies that the time derivative of the scalar field today (with respect to t_*) must satisfy

$$A_0^{-1} \left| \alpha_0 + \frac{\beta_0 \alpha_0}{(1 + \alpha_0^2)} \right| \left| \frac{d\varphi_{*0}}{dt_*} \right| < 3 \times 10^{-12} \text{ years}^{-1}. \quad (10.27)$$

Unfortunately, the experimental limit (10.20) on the time variation of the gravitational constant does not imply any constraint on $2\dot{A}/A = -\dot{F}/F$. Indeed, G_{cav} can be almost constant even if A varies a lot. An example is that of the Barker theory [7] in which $A = \cos \varphi_*$ so that $G_{\text{N}} = G_{\text{cav}} = G_*$ is strictly constant independently of the time variations of $A[\varphi_*(t)]$.

10.1.3 Cosmological aspects

Despite the above-mentioned strong local constraints, deviations from general relativity are far less constrained on cosmological scales. We now expand on the cosmology of the scalar-tensor theory [5, 8, 9].

10.1.3.1 Friedmann equations

We consider a Friedmann–Lemaître space-time with metric (3.1). We distinguish between the form of this metric in the Jordan and Einstein frames and denote by a and a_* , t and t_* , η and η_* , respectively, the scale factors, proper times, and conformal

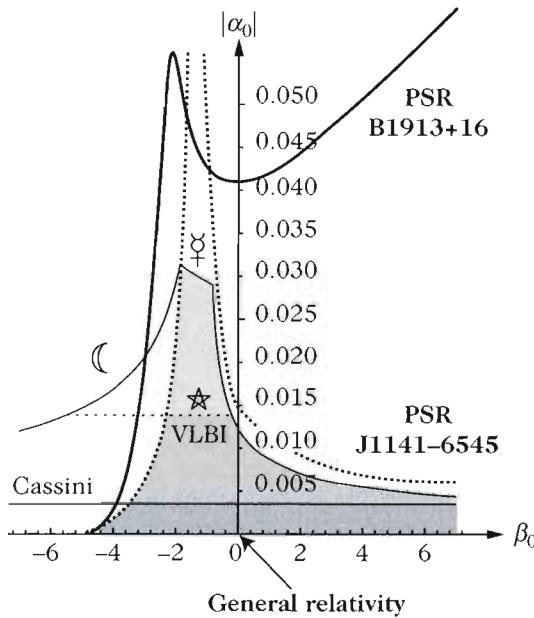


Fig. 10.1 Constraints on the parameters (α_0, β_0) of a scalar-tensor theory with coupling expanded as $\ln A = \alpha_0(\varphi_* - \varphi_0) + \frac{1}{2}\beta_0(\varphi_* - \varphi_0)^2 + \mathcal{O}[(\varphi_* - \varphi_0)^3]$ obtained from the constraints in the Solar System (see Fig. 1.9, same labels) and the two binary pulsars, PSR B1913+16 and PSR J1141-6545. The shaded region is the allowed region; the vertical axis ($\beta_0 = 0$) corresponds to the Brans–Dicke theory. From Ref. [6].

times in each of these two frames. The transformation (10.8) implies that the metrics are related by $ds^2 = A^2(\varphi_*)ds_*^2$ so that

$$dt = A(\varphi_*)dt_*, \quad d\eta = d\eta_*, \quad a = A(\varphi_*)a_*. \quad (10.28)$$

Note that the comoving spatial coordinates remain unchanged $|dx| = |dx_*|$. We find that the redshifts in both frames are related by

$$1 + z = \frac{A(\varphi_{*0})}{A(\varphi_*)}(1 + z_*), \quad (10.29)$$

and that the physical lengths associated with the same comoving length are related by

$$\ell_*^{\text{phys}} = \frac{A(\varphi_{*0})}{A(\varphi_*)}\ell^{\text{phys}}. \quad (10.30)$$

The energy-momentum tensors are then expressed in the form (3.24) where $u^\mu = dx^\mu/|ds|$ and $u_*^\mu = dx_*^\mu/|ds_*|$. The relation between $T_{\mu\nu}$ and $T_{\mu\nu}^*$ implies that

$$\rho_* := A^4\rho, \quad P_* = A^4P. \quad (10.31)$$

In the Jordan frame, (10.4) and (10.6) then take the form

$$3F \left(H^2 + \frac{K}{a^2} \right) = \kappa_* \rho + \frac{Z}{2} \dot{\varphi}^2 - 3H\dot{F} + U, \quad (10.32)$$

$$-2F \left(\dot{H} - \frac{K}{a^2} \right) = \kappa_* (\rho + P) + Z\dot{\varphi}^2 + \ddot{F} - H\dot{F}, \quad (10.33)$$

$$Z(\dot{\varphi} + 3H\dot{\varphi}) = 3F_{,\varphi} \left(\dot{H} + 2H^2 + \frac{K}{a^2} \right) - \frac{1}{2} Z_{,\varphi} \dot{\varphi}^2 - U_{,\varphi}, \quad (10.34)$$

$$\dot{\rho} + 3H(1+w)\rho = 0. \quad (10.35)$$

The Hubble constant is defined by $H = \dot{a}/a$ and all the time derivatives are with respect to t . Equation (10.35) is, as expected, identical to the one obtained in the standard case (3.27).

The corresponding equations in the Einstein frame are very similar to those in general relativity,

$$3 \left(H_*^2 + \frac{K}{a_*^2} \right) = \kappa_* \rho_* + \dot{\varphi}_*^2 + 2V(\varphi_*), \quad (10.36)$$

$$\frac{\ddot{a}_*}{a_*} = -\frac{\kappa_*}{6}(\rho_* + 3P_*) - \frac{2}{3}\dot{\varphi}_*^2 + \frac{4}{3}V, \quad (10.37)$$

$$\ddot{\varphi}_* + 3H_*\dot{\varphi}_* = -V_{,\varphi_*} - \frac{1}{2}\kappa_*\alpha(\varphi_*)(\rho_* - 3P_*), \quad (10.38)$$

$$\dot{\rho}_* + 3H_*(1+w)\rho_* = \alpha(\varphi_*)(\rho_* - 3P_*)\dot{\varphi}_*, \quad (10.39)$$

where the time derivatives are now taken with respect to t_* . In this frame, the evolution of the density is more complicated and is given by $\rho_* \propto A^{1-3w}a_*^{-3(1+w)}$ for a fluid with constant equation of state w . The analogues of (10.32)–(10.39) in conformal time are easily obtained [8].

Note that if the field potential vanishes, $V = 0$, then (10.37) implies that $\ddot{a}_* < 0$, so that in the Einstein frame, the expansion of the Universe is decelerating. However, due to (10.28), the observed expansion, i.e. that in the Jordan frame, can be accelerated, i.e. \ddot{a} may be positive. The constants H and H_* are related by

$$H = A^{-1} \left[H_* + \alpha(\varphi_*) \frac{d\varphi_*}{dt_*} \right],$$

from which one can directly deduce the (observed) expansion rate of the Universe (i.e. in the Jordan frame) once the Friedmann equations are solved in the Einstein frame.

10.1.3.2 Attractor mechanism towards general relativity

Many scalar-tensor models are naturally attracted towards general relativity during the cosmological evolution [10]. To understand this mechanism, let us rewrite the Klein–Gordon equation (10.38) in the Einstein frame, using $p = \ln a_*$, i.e. the number of e-folds in the Einstein frame, as the time variable. Defining

$$v(\varphi_*) = 2 \frac{V(\varphi_*)}{\kappa_* \rho_*}, \quad (10.40)$$

and assuming a matter content with a constant equation of state w , the Klein-Gordon equation takes the form

$$\frac{2[1+v(\varphi_*)]}{3-\varphi_*'^2}\varphi_*''+[1-w+2v(\varphi_*)]\varphi_*'= -\alpha(\varphi_*)(1-3w)-v(\varphi_*)\frac{d\ln V}{d\varphi_*}, \quad (10.41)$$

where a prime stands for a derivative with respect to p .

Another way to write this equation is to separate the potential and kinetic parts of the scalar field, and to set $\rho_V = -P_V = V/4\pi G_*$, $\rho_T = \rho + \rho_V$ and $P_T = P + P_V$. Equation (10.38) then takes the form

$$\frac{2}{3-\varphi_*'^2}\varphi_*''+\left(1-\frac{P_T}{\rho_T}\right)\varphi_*'= -\alpha\frac{\rho-3P}{\rho_T}-\alpha_V\frac{\rho_V-3P_V}{\rho_T},$$

with $\alpha_V = V'/4V$.

To illustrate the attraction mechanism at work, let us consider a model for which the coupling function is of the form $A = \exp[b(\varphi_*)]$ with $b(\varphi_*) = \frac{1}{2}\beta\varphi_*^2$ and β constant, so that $\alpha(\varphi_*) = \beta\varphi_*$. Equation (10.40) takes the form of the equation of motion of a relativistic particle with a velocity-dependent mass, $m(\varphi_*) = 2/(3-\varphi_*'^2)$, subject to a damping force $-(1-w)\varphi_*'$, and evolving in a potential $(1-3w)b(\varphi_*)$,

$$\frac{2}{3-\varphi_*'^2}\varphi_*''+(1-w)\varphi_*'= -\beta(1-3w)\varphi_*. \quad (10.42)$$

The local positivity of energy imposes $m(\varphi_*) > 0$ so that $\varphi_*'^2 < 3$. During the radiation era, $w = \frac{1}{3}$ and the scalar field decouples from matter. Thus, it evolves freely and freezes to a constant value, $\varphi_{*,R}$, independently of its initial velocity.

Equation (10.41) can be solved analytically during the radiation era when the potential is negligible to obtain

$$\varphi_* = \varphi_i - \sqrt{3} \left[\frac{\alpha_i e^{-(p-p_i)} + \sqrt{1 + \alpha_i^2 e^{-2(p-p_i)}}}{\alpha_i + \sqrt{1 + \alpha_i^2}} \right],$$

where $\varphi_i = \varphi_*(p_i)$ and $\alpha_i\sqrt{3} = \varphi_i'/\sqrt{1-\varphi_i'^2/3}$. Thus, as long as $\varphi_i' \ll \sqrt{3}$ the variation of φ_* during the radiation era is $\Delta\varphi_* \sim \varphi_i'$ and the field is at rest after a few e-folds. In the matter era, the evolution of φ_* is that of a damped oscillator with vanishing initial speed. So φ_* evolves towards the minimum of the coupling function where $\alpha = 0$, if $\beta > 0$. The scalar-tensor theory then evolves in such a way as to end up infinitely close to general relativity today.

10.1.3.3 Cosmological constraints

From a cosmological viewpoint, there are currently few constraints on scalar-tensor theories. The observation of the cosmic microwave background [9] can highlight some effects, but they are often degenerate with other cosmological parameters. The study of weak gravitational lensing effects, combined with the cosmic microwave background, is a promising tool [8].

The only definite constraints are obtained from primordial nucleosynthesis (see Chapter 4). The analysis of helium-4 abundance indicates that the number of relativistic degrees of freedom cannot vary by more than 20% with respect to its nominal value $g_r = 10.75$.

For a Universe with Euclidean spatial sections, the Friedmann equation (10.32) implies

$$\mathcal{H}^2 = \kappa_0 \frac{\pi^2}{90} a^2 g_r \left(1 + \frac{\delta g_r}{g_r} \right) T^4, \quad (10.43)$$

with $\delta g_r/g_r = F_0/F(z_{\text{nuc}}) - 1$, neglecting the contribution of the scalar field to the energy density. The helium-4 constraint imposes that $\delta g_r/g_r$ has to be lower than 20%, and we thus obtain the rough constraint

$$0.8 \leq \left| \frac{F_0}{F_{\text{nuc}}} \right| = \left| \frac{A_{\text{nuc}}^2}{A_0^2} \right| \leq 1.2. \quad (10.44)$$

This estimate is very general. As soon as one specifies the model, one can put stronger constraints through a detailed analysis (see, e.g., Ref. [11]).

In brief, cosmology can constrain these models in three regions: $z \sim 10^8$ with nucleosynthesis, $z \sim 10^3$ with the cosmic microwave background and $z \sim 1$ with lensing effects.

10.1.4 Phenomenological aspects

The attractor mechanism opens up interesting possibilities for phenomenological cosmology. However, since each model involves two arbitrary functions, this can easily accommodate a large number of possibilities. The feasibility of reconstructing these functions from cosmological observations was detailed in Ref. [5].

Here we shall only discuss a few illustrative examples concerning the evolution of large-scale structures and the relation with the recent acceleration of the Universe.

10.1.4.1 Influence on the evolution of perturbations

Scalar-tensor theories can be studied along the lines developed in Chapter 5. For definiteness, let us consider the modifications that we would obtain by working in the Jordan frame. In this case, the fluid equations of evolution are not modified and only the Einstein equations (5.117)–(5.120) take a new form.

If $\delta\varphi$ is the perturbation of the scalar field in Newtonian gauge, setting $\delta F \equiv F_{,\varphi}\delta\varphi$, the Einstein field equations take the form

$$\Psi - \Phi = \kappa_* a^2 P \bar{\pi} + \frac{\delta F}{F}, \quad (10.45)$$

$$2F(\Psi' + \mathcal{H}\Phi) + F'\Phi = -\kappa_* a^2 \rho(1+w)V + Z\varphi'\delta\varphi + \delta F' - \mathcal{H}\delta F, \quad (10.46)$$

$$\begin{aligned} 2F\Delta\Phi - (Z\varphi'^2 - 3\mathcal{H}F')\Phi &= 3F'\Phi' + \kappa_* \rho a^2 \delta^C - \kappa_* P \Delta \bar{\pi} + \frac{1}{2} \varphi'^2 Z_{,\varphi} \delta\varphi \\ &\quad - \left[\Delta + 3 \left(\mathcal{H}^2 + \frac{F'^2}{F^2} \right) \right] \delta F + 3 \frac{F'}{F} \delta F' \\ &\quad + (U_{,\varphi} a^2 + 3\mathcal{H}Z\varphi')\delta\varphi + Z\varphi'\delta\varphi'. \end{aligned} \quad (10.47)$$

The perturbed Klein–Gordon equation takes a more compact form if we restrict ourselves to $Z = 1$ (see Ref. [5] for the general form with $Z \neq 1$),

$$\begin{aligned} \delta\varphi'' + 2\mathcal{H}\delta\varphi' - [\Delta + 3\mathcal{H}'F_{,\varphi\varphi} - U_{,\varphi\varphi}a^2] \delta\varphi &= (\Phi' + 3\Psi')\delta\varphi' \\ - 2a^2\Phi U_{,\varphi} - [\Delta(\Phi - 2\Psi) + 3(\Psi'' + 3\Psi' + \mathcal{H}\Phi')] F_{,\varphi}. \end{aligned} \quad (10.48)$$

A simple application is obtained by restricting ourselves to the Newtonian regime in the matter-dominated era. In this case, the fluid equations of evolution (see Chapter 5) take the form $\delta'_m = -\Delta V + 3\Psi'$ and $V' + \mathcal{H}V = -\Phi$, while the Poisson equation (10.47) becomes

$$F\Delta\Phi = 4\pi G_*\rho a^2\delta_m - \frac{F_{,\varphi}}{2}\Delta\delta\varphi, \quad (10.49)$$

and the perturbed Klein–Gordon equation (10.48)

$$(\Delta - U_{,\varphi\varphi}a^2)\delta\varphi = F_{,\varphi}\Delta(\Phi - 2\Psi). \quad (10.50)$$

So, in the linear regime, if the field is light, i.e. if $U_{,\varphi\varphi}$ remains small compared to the wavelength of the considered modes, then (10.45) and (10.50) imply that

$$\delta\varphi \simeq -\frac{FF_{,\varphi}}{F + 2F_{,\varphi}^2}\Phi, \quad (10.51)$$

and the Poisson equation (10.49) can be rewritten as

$$\Delta\Phi \simeq 4\pi G_{\text{cav}}\rho a^2\delta_m. \quad (10.52)$$

Finally, the evolution equation for matter perturbations takes the form

$$\delta''_m + \mathcal{H}\delta'_m - 4\pi G_{\text{cav}}\rho a^2\delta_m = 0. \quad (10.53)$$

This illustrates the influence from modifications of the theory of gravity. A detailed discussion of the various models can be found in Refs. [8, 9]. Since G_{cav} can depend on time and relax toward G_N , the growth rate of density perturbation will be modified.

10.1.4.2 Quadratic coupling

A simple model [10] is provided by a quadratic coupling function, namely with $b(\varphi_*) = a_m + \frac{1}{2}\beta(\varphi_* - \varphi_m)^2$, where φ_m is the value of the field at the minimum of the coupling. Without loss of generality, a field redefinition leads to

$$b(\varphi_*) = \frac{1}{2}\beta\varphi_*^2, \quad (10.54)$$

so that $\alpha_0 = \beta\varphi_0^*$ and $\beta_0 = \beta$.

The field evolution can be obtained analytically if the field is massless. Indeed, during the matter-dominated era, $\varphi_*'^2 \ll 3$, and the Klein–Gordon equation simplifies to

$$\frac{2}{3}\varphi_*'' + \varphi_*' + \beta\varphi_* = 0. \quad (10.55)$$

Depending on the value of β , we obtain two different regimes. If $\beta < 3/8$, for weak coupling, we have

$$\varphi_*(z_*) = A_+(1+z_*)^{3(1+r)/4} + A_-(1+z_*)^{3(1-r)/4} \quad (10.56)$$

with $r = \sqrt{1 - 8\beta/3}$. If $\beta > 3/8$, the field undergoes damped oscillations,

$$\varphi_*(z_*) = (1+z_*)^{3/4} \left\{ A \cos \left[\frac{3}{4}r \ln(1+z_*) \right] + B \sin \left[\frac{3}{4}r \ln(1+z_*) \right] \right\}, \quad (10.57)$$

so that G_{cav} can oscillate. During the radiation era, the solution for φ_* is

$$\varphi_* = \pm \sqrt{3} \operatorname{arctanh} \sqrt{1 - A(1+z)^{-2}} + B. \quad (10.58)$$

These analytical behaviours correctly reproduce the numerical solution depicted in Fig. 10.2.

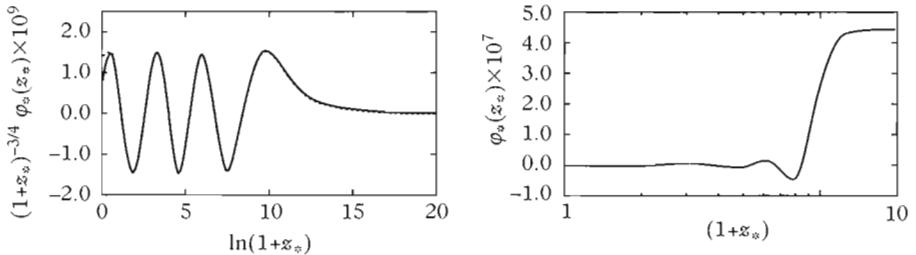


Fig. 10.2 Evolution of a massless scalar field in the Einstein frame in the case of a quadratic coupling theory $\beta = 10^2/8\pi$. From Ref. [8].

In a Universe filled with a matter–radiation mixture, (10.41) takes the form

$$y(y+1) \frac{d^2\varphi_*}{dy^2} + \frac{1}{2}(5y+4) \frac{d\varphi_*}{dy} + \frac{3}{2}\beta\varphi_* = 0,$$

with $y = a_*/a_{\text{eq}*}$, as long as the potential is negligible. This equation has a solution in terms of hypergeometric functions $\varphi_*(y) = \varphi_{*,R} F(v, v^*, 2; y)$ with $v = \frac{3}{4} - i\sqrt{\frac{3}{2}(\beta - \frac{3}{8})}$, which makes it possible to relate the value of the scalar field to its initial value in the radiation era.

Primordial nucleosynthesis has been studied in this model [11, 12] and found to impose strong constraints on the set of parameters (α_0, β) , which are summarized in Fig. 10.3. Despite the strong constraints on α_0 imposed by the Solar System, some

models enjoying strong values of β were admissible. In this precise case, nucleosynthesis imposes the constraints

$$\beta \alpha_0^2 < 10^{-6.5} \left(\frac{\Omega_m h^2}{0.15} \right)^{-3/2}, \quad \beta \gtrsim 0.5,$$

assuming the Universe has Euclidean spatial sections. This shows the complementarity of cosmological and local constraints.

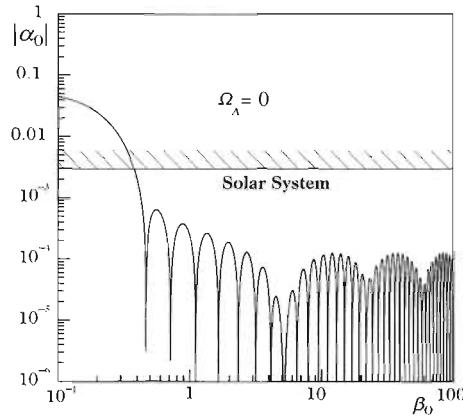


Fig. 10.3 Maximal values of α_0 as a function of $\beta_0 = \beta$ for a scalar-tensor theory with quadratic coupling (10.54) assuming a vanishing cosmological constant. The dotted line represents the Solar-System constraints. From Ref. [12].

10.1.4.3 Introducing a cosmological constant

In the previous model, the dilaton was assumed to be massless since it has no potential. Therefore, it cannot play the role of a quintessence-like field and explain the recent acceleration of the Universe.

It is interesting to extend it to models containing a cosmological constant [12]. In general relativity, such a constant amounts to adding a constant potential to the scalar field such that the latter does not change mass. The situation is more complex in scalar-tensor theory and there is no unique way to generalize the Λ CDM or quintessence model (see Section 12.2 of Chapter 12).

The first solution is to add a constant to the potential U in the Jordan frame. In this case, this constant corresponds to a constant energy density and is thus associated to a fluid with equation of state $w = -1$, which seems a good generalization of the cosmological constant. However, (10.11) implies that the true spin-0 degree of freedom, φ_* , evolves in a potential $V = \frac{1}{2}\Lambda A^4$. So it acquires a mass.

Assuming a constant potential in the Einstein frame, φ_* remains massless, but this constant corresponds to an energy density $\rho = 2\Lambda F^2/\kappa_*$, which is thus not constant.

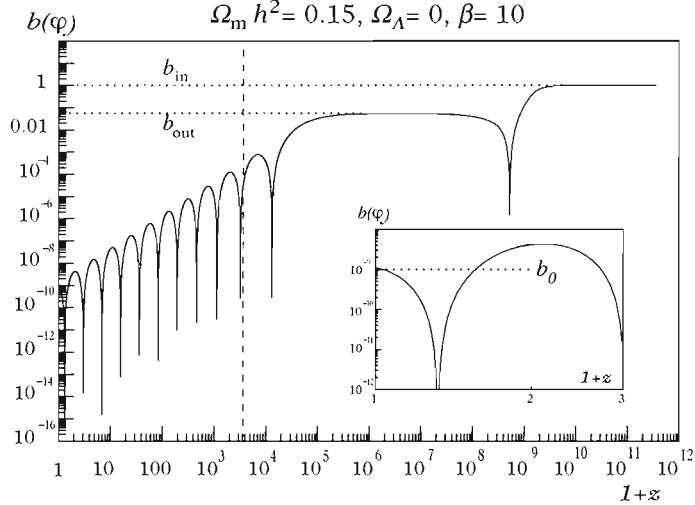


Fig. 10.4 Evolution of $b(\varphi_*)$ as a function of redshift for $\beta = 10$ assuming that $b \rightarrow b_{\text{in}} = 1$ when $z \rightarrow \infty$. The vertical dashed line corresponds to the matter–radiation equality. During the radiation era the scalar field freezes to a constant value b_{out} . It can, however, oscillate during mass thresholds, i.e. electron–positron annihilation. In the matter era the field oscillates, while being damped as $(1+z)^{3/4}$. From Ref. [12].

10.1.4.4 Runaway dilaton model

In the context of quintessence (see Section 12.2), an extended class of quintessence models have been introduced [13]. In this case, the scalar field of the scalar-tensor theory also plays the role of a quintessence field. Models of this type, inspired by string theory, have been proposed [14, 15]; they are called *runaway dilaton* models, the dilaton being the name given to the scalar field in string theory.

Interestingly, the attractor mechanism thanks to which the scalar-tensor theory can evolve cosmologically towards general relativity, discussed earlier, can be generalized [16] when the field evolves in a potential of the form

$$V(\varphi_*) = M^4 \varphi_*^{-m}. \quad (10.59)$$

For this, it is sufficient that the coupling to matter be of the form

$$\alpha(\varphi_*) = -B e^{-\beta \varphi_*}. \quad (10.60)$$

This mechanism will be effective as soon as the coupling and the potential are proportional, $A \propto V$, since we then have $\alpha \propto \alpha_V$.

The scalar field is then naturally driven towards $\varphi_* \rightarrow \infty$ during the cosmological evolution. During the radiation-dominated era, the coupling is not efficient since the field starts slowly rolling only in the matter era. Assuming that it is still in the slow-roll regime today, the Klein–Gordon equation then takes the form

$$\varphi'_* \simeq Be^{-\beta\varphi_*}, \quad (10.61)$$

for which an approximate solution is

$$e^{\beta\varphi_*} = e^{\beta\varphi_0} + B\beta \ln(1 + z_*). \quad (10.62)$$

Today, $\alpha_0 = -Be^{-\beta\varphi_0}$ and $\beta_0 = -\beta\alpha_0$ and the rolling of the field can explain the present acceleration of the Universe.

10.1.5 $f(R)$ gravity and scalar-tensor theory

In the previous analysis, just as in general relativity, gravity is described by the Einstein–Hilbert action. Nevertheless, one can conceive metric theories of gravity for which the Einstein–Hilbert Lagrangian is another, arbitrary, function of the curvature scalar,

$$S = \frac{1}{16\pi G_*} \int f(R) \sqrt{-g} d^4x + S_m[g_{\mu\nu}; \text{matter}]. \quad (10.63)$$

Such a theory leads to the (modified) Einstein equations

$$f'(R)R_{\mu\nu} - \frac{1}{2}f(R)g_{\mu\nu} - \nabla_\mu \partial_\nu f'(R) + g_{\mu\nu} \square f'(R) = \kappa T_{\mu\nu}, \quad (10.64)$$

where a prime indicates a derivative of the function with respect to its argument, i.e. $f'(R) \equiv df/dR$. Interestingly, one can show that all these classes of theories reduce to a scalar-tensor theory [17, 18]. To see this, let us introduce an auxiliary field φ and let us consider the action

$$S = \frac{1}{16\pi G_*} \int [f'(\varphi)R + f(\varphi) - \varphi f'(\varphi)] \sqrt{-g} d^4x + S_m[g_{\mu\nu}; \text{matter}]. \quad (10.65)$$

The variation of this action with respect to this scalar field indeed implies, if $f''(\varphi) \neq 0$ (a case that is equivalent to general relativity with a cosmological constant), that

$$R - \varphi = 0. \quad (10.66)$$

This constraint permits us to conclude that the Einstein equation derived from (10.65), namely

$$f'(\varphi)G_{\mu\nu} - \nabla_\mu \partial_\nu f'(\varphi) + g_{\mu\nu} \square f'(\varphi) + \frac{1}{2}[\varphi f'(\varphi) - f(\varphi)]g_{\mu\nu} = \kappa T_{\mu\nu}, \quad (10.67)$$

reduces to (10.64). Note that, even if the action (10.65) does not possess a kinetic term for the scalar field, the theory is well defined since the true spin-0 degree of freedom clearly appears in the Einstein frame, and with a positive energy.

The change of variable (10.9) implies that we can choose

$$\varphi_* = \frac{\sqrt{3}}{2} \ln f'(\varphi), \quad (10.68)$$

so that the theory in the Einstein frame is defined by

$$A^2 \propto \exp[-4\varphi_*/\sqrt{3}], \quad V = \frac{1}{4} \left\{ \varphi(\varphi_*) e^{2\varphi_*/\sqrt{3}} - f[\varphi(\varphi_*)] \right\} e^{-4\varphi_*/\sqrt{3}}. \quad (10.69)$$

This example highlights the importance of looking for the true degrees of freedom of the theory. A field redefinition can be a useful tool to show that two theories are actually equivalent.

10.2 Quantum field theory in curved space-time

The standard model of the three non-gravitational interactions has been developed in Minkowski space-time. During the majority of the history of the Universe, space-time can be described classically. The first extension necessary to describe the primordial phases is then to consider the extension of the field theory presented in Chapter 2 to a curved space-time [19]. We have used some results of this theory in Chapter 8.

10.2.1 Quantum physics, classical gravity

The entire aim of quantum field theory in curved space-time amounts to treating gravity with general relativity, i.e. classically, whereas fields are governed by the laws of quantum mechanics.

10.2.1.1 Energy scales

In most situations, it is not necessary to take into account effects from quantum gravity, no matter which theory describes it, mainly because these effects are supposed to appear at scales of the order of the Planck mass.

This can be seen heuristically in the following way. Let us compare the Newtonian gravitational potential between two masses m_1 and m_2 , $V_{\text{grav}} = G_N m_1 m_2 / r$ to the Coulomb potential $V_{\text{elec}} = \alpha q_1 q_2 / r$ between two charges q_1 and q_2 , where α is dimensionless. The coupling between two massive particles is written just as in electromagnetism: $V_{\text{grav}} = G_N E_{\text{cm}}^2 \tilde{m}_1 \tilde{m}_2 / r$, with $\tilde{m}_i = m_i / E_{\text{cm}}$ ($i = 1, 2$) two dimensionless gravitational ‘charges’, such that the effective gravitational coupling constant actually varies with the square of the interaction energy. This variation is very similar to the variation of α with energy, only the latter is logarithmic. Effects from quantum gravity finally arise when the coupling constant is of order unity, and thus $G_N E_{\text{cm}}^2 \sim 1$, i.e. when the energy scale of the interaction is of the order of the Planck mass, $E_{\text{cm}} \sim M_P$.

In most situations, we are interested in energy scales of the order of E_{cut} at best, that is three orders of magnitudes below M_P . It is therefore perfectly justified, for most applications, to treat space-time classically, while particles are subject to quantum effects.

10.2.1.2 Energy renormalization

Let us consider (2.89)

$$\hat{H}(t) = \frac{1}{2} \int d^3k \omega_k (\hat{a}_k^\dagger \hat{a}_k + \hat{a}_k \hat{a}_k^\dagger) = \int d^3k \omega_k \left[\hat{\mathcal{N}}_k + \frac{1}{2} \delta^{(3)}(\mathbf{0}) \right], \quad (10.70)$$

giving the value of the total energy of a scalar field as a function of the creation and annihilation operators. In ordinary field theory, we argued that the zero-point energy, appearing in (10.70) in the form of a factor $\delta^{(3)}(\mathbf{0})$, did not have any physical meaning, and that we could therefore ‘renormalize’ the energy at will. A way to proceed is to notice that there is always an ambiguity in the definition of the operator products, and that a valid choice is to take the normal product for which all the creation operators are put on the left and the annihilation operators on the right. With the notation ‘ $\langle : \hat{H} :\rangle$ ’ to define the normal-ordered operator, (10.70) becomes

$$\langle : \hat{H}(t) : \rangle = \int d^3k \omega_k \hat{a}_k^\dagger \hat{a}_k = \int d^3k \omega_k \hat{\mathcal{N}}_k, \quad (10.71)$$

and the embarrassing term is automatically suppressed.

As soon as we couple the scalar field to gravity, such a renormalization is no longer completely innocent: since gravity couples to every form of energy, it is no longer possible to modify the value of the total energy, not even by a constant. The cosmological constant problem partly arises from this difficulty.

10.2.1.3 Semi-classical Einstein equations

Since quantum effects from the gravitational field are not important during the majority of the history of the Universe, we can describe a physical state as one classical component, the geometry, and one quantum component, that is a normalized state $|\psi\rangle$. The geometric part of Einstein equations is thus a trivial operator on this state, and we have $\langle\psi|G_{\mu\nu}|\psi\rangle = G_{\mu\nu}\langle\psi|\psi\rangle = G_{\mu\nu}$, and, if the Universe is in the quantum state $|\psi\rangle$, at the level of particle physics, we have

$$G_{\mu\nu} = \kappa \langle\psi|\hat{T}_{\mu\nu}|\psi\rangle, \quad (10.72)$$

where we can see that the expectation value of the energy-momentum operator $\hat{T}_{\mu\nu}$ plays the role of the source.

The elementary renormalization of the operator $\hat{T}_{\mu\nu}$ made earlier thanks to the normal ordering of the creation and annihilation operators, which is possible in Minkowski space-time, must be done in a more subtle way [20] for a more general space. It can lead to so-called backreaction effects in cosmology, and can modify the dynamics of the scale factor [21]. These effects, which are beyond the scope of this book, will probably be considered seriously in the future. Furthermore, it leads to the prediction of the existence of a cosmological constant discussed in detail in Chapter 12. The contribution (10.70) is indeed of the order of E_c^4 without performing the renormalization (10.71), E_c being a cutoff energy assumed to be of the order of M_p .

10.2.1.4 *D*-dimensional scalar field

As in Chapter 2, the typical theory we are interested in here is that of a scalar field, and the action we will consider is the generalization of the action (1.80) with no cosmological constant

$$\mathcal{S} = -\frac{1}{2} \int d^D x \sqrt{|g|} [\partial_\mu \phi \partial^\mu \phi + 2V(\phi) - \xi \phi^2 R], \quad (10.73)$$

in a *D*-dimensional space-time. This choice of an arbitrary number of dimensions will allow us to use results following from string theory, with more than 4 dimensions (or yet with only two dimensions as in the case of the example of Section 10.2.3). The last term in (10.73) allows for a non-minimal coupling between gravity and the scalar field (see also Section 10.1.1).

The Klein–Gordon equation for the scalar field is modified to

$$\left(\frac{1}{\sqrt{|g|}} \partial_\mu \sqrt{|g|} \partial^\mu + \xi R \right) \phi = \frac{dV}{d\phi}. \quad (10.74)$$

The so-called minimal-coupling amounts to choosing $\xi = 0$. This choice is often made to simplify matters, but in any case if the theory (10.73) was verified experimentally, the parameter ξ would be a physically measurable parameter.

10.2.1.5 Conformal invariance

On performing a metric transformation of the form

$$\tilde{g}_{\mu\nu} = \Omega^2(x) g_{\mu\nu}, \quad \text{and} \quad \tilde{\phi} = \Omega^{(2-D)/2} \phi, \quad (10.75)$$

called a *conformal transformation*, the scalar curvature transforms as

$$\tilde{R} = \frac{R}{\Omega^2} - 2 \frac{D-1}{\Omega^3} g^{\mu\nu} \nabla_\mu \nabla_\nu \Omega - \frac{(D-1)(D-4)}{\Omega^4} g^{\mu\nu} \nabla_\mu \Omega \nabla_\nu \Omega, \quad (10.76)$$

such that the action is invariant, up to a total derivative, if $V(\phi) = 0$ (massless scalar field with no interaction), and as long as the arbitrary coefficient ξ is

$$\xi = \xi_{\text{conformal}} = \frac{D-2}{4(D-1)}.$$

The theory is then conformally invariant. This type of transformation is interesting in cosmology as the Friedmann–Lemaître metric with flat spatial sections (a good approximation at high energy/redshift) happens to be conformally flat, in other words we can write it in the form $g_{\mu\nu} = a^2(\eta) \eta_{\mu\nu}$, which explains the nomenclature of η as conformal time.

If $D = 2$, we see that the theory is conformally invariant as long as $\xi = 0$: thus, minimal and conformal coupling are the equivalent in two dimensions. In four dimensions, the conformal coupling corresponds to $\xi = \frac{1}{6}$.

10.2.2 Particle creation

If quantum fields can modify the space-time dynamics, they can in turn also act as a classical source and produce particles. When cosmological perturbations themselves are considered as quantum fields, the time-evolution of the scalar field can be used to create excitations and be seen as the source for the generation of primordial fluctuations. This can also be used to fix the initial conditions of these perturbations since one can ask, for instance, that the Universe be in its vacuum state at the initial time, corresponding in principle to a completely defined state.

10.2.2.1 Number of particles

Let us consider the mode decomposition of (2.86) with the modes (2.85). The latter are essentially defined as eigenfunctions of the vector $\partial/\partial t$, which is a Killing vector, in the Minkowski case. The eigenvalues are $-i\omega_k$, defining the states of positive frequency $\omega_k > 0$. The vacuum, defined by (2.90), is invariant under transformations of the Poincaré group.

The situation becomes more complicated as soon as transformations that are not in the Poincaré group are considered, while remaining in the context of special relativity. For instance, we can place ourselves in an accelerating frame with respect to that in which the previous modes are defined. We then see that the previous vacuum state is not invariant: the particle number operator \bar{N} defined by the new mode basis in the accelerated frame and applied on the vacuum state $|0\rangle$ of the first frame, is no longer of zero eigenvalue, $\bar{N}|0\rangle \neq 0$, which is interpreted as saying that the number of particles depends on the frame in which it is measured. The concept of particles has therefore no intrinsic meaning.

The number of particles is actually only well defined in an inertial frame, and this could even be an alternative definition of such a reference frame. Notice, moreover, that a detector in free fall is a particular case, and according to the equivalence principle, it does not measure the same number of particles as another non-inertial detector, in acceleration.

The problem is related to the fact that the basis of the eigenmodes on which the fields are expanded, is composed of global quantities, i.e. defined on the entire space-time (or at least on macroscopic subspaces). As a consequence, a given decomposition depends on the entire past history of the observer. It is, by the way, a possible argument in favour of the semi-classical theory (10.72) since only local terms are involved in this equation. If $\langle\psi|\hat{T}_{\mu\nu}|\psi\rangle = 0$ for a given observer, it remains zero in all possible frames since it must transform as a tensor.

10.2.2.2 Eigenmodes in curved space-time

In the case where the metric is no longer Minkowski, the situation is even worse than in an accelerated reference frame. The entire previous section becomes dubious, there is in particular no reason why the vector ∂_t should remain a Killing vector. In other words, the eigenmode basis is, in the curved case, a basis of functions $u_i(x)$. The index i represents collectively a set of indices required to identify the modes. It can be continuous, discrete, or a mixture of both cases. These modes satisfy (10.74)

with $V = \frac{1}{2}m^2\phi^2$ so as to remain linear in the field. The flat-space scalar product is generalized to

$$\langle f, g \rangle \equiv -i \int_{\Sigma} f \overleftrightarrow{\partial}_{\mu} g^* \sqrt{|g_{\Sigma}|} u^{\mu} d\Sigma, \quad (10.77)$$

where the derivative on the right and left is defined by (2.69), where Σ is a space-like hypersurface with normal vector u^{μ} (hence time-like) directed towards the future, and $d\Sigma$ the volume element of Σ . These modes form an orthonormal basis if we require, setting $\delta(i-j)$ to indicate either a Kronecker or Dirac ' δ ' depending on whether the index is integer or real,

$$\langle u_i, u_j \rangle = \delta(i-j) = -\langle u_i^*, u_j^* \rangle, \quad \text{and} \quad \langle u_i, u_j^* \rangle = 0, \quad (10.78)$$

which is always possible to impose. It is indeed the generalization of (2.87) for a curved space. The functions (2.85) satisfy

$$\langle u_k, u_p \rangle = -i \int d^3x u_k \overleftrightarrow{\partial}_t u_p^* = \delta^{(3)}(\mathbf{k} - \mathbf{p}) = -\langle u_k^*, u_p^* \rangle, \quad \text{and} \quad \langle u_k, u_p^* \rangle = 0.$$

In a space-time that is not Minkowski, we do not know how to treat the state space in a general way. We thus proceed in a manner analogous to the theory of interacting fields, by considering that there exists some space-like (in the case of static configurations such as black holes for which we compute the Hawking radiation [22]) or time-like (in cosmology) regions, in which the space-time is asymptotically Minkowski to a good approximation. In this case, the vacuum is well defined in these regions, and it is sufficient to compute the overlap of the vacuum states to know the quantity of particles created by passing from one to another. We will see an application of this mechanism in Section 10.2.3.

10.2.2.3 Cosmological modes

There is another case for which the vacuum state can be defined in a consistent way: it is the Friedmann–Lemaître solution, and in particular for the de Sitter case. Consider the metric (3.3) with $K = 0$ (to simplify, the general case can be treated in a similar way).

The mode equation (10.74) for $D = 4$ and $V = \frac{1}{2}m^2\phi^2$ becomes

$$u_k'' + \left\{ k^2 + a^2 \left[m^2 - \left(\xi - \frac{1}{6} \right) R \right] \right\} u_k = 0, \quad (10.79)$$

where we have considered the expansion in the eigenvectors of the Laplacian [see (B.56)], where the field is from now on quantum

$$\hat{\phi}(\mathbf{x}, \eta) = \frac{1}{a} \int \frac{d^3k}{(2\pi)^{3/2}} \left[\hat{a}_k Q_k(\mathbf{x}) u_k(\eta) + \hat{a}_k^\dagger Q_k^*(\mathbf{x}) u_k^*(\eta) \right], \quad (10.80)$$

defining the creation and annihilation operators, and the scalar curvature is $R = -6a''/a$. We may notice that (10.79) is valid independently of the time-dependence of the scale factor.

The field conjugate momentum is

$$\hat{\pi}(x) = \frac{\partial \mathcal{L}}{\partial \dot{\phi}'(x)} = a^2 \hat{\phi}'(x),$$

where a prime indicates a derivative with respect to the conformal time η . To satisfy the canonical commutation relations at equal time (2.84), namely here $[\hat{\phi}(\mathbf{x}, \eta), \hat{\pi}(\mathbf{y}, \eta)] = i\delta^{(3)}(\mathbf{x} - \mathbf{y})$ and also requiring (2.83) between \hat{a}_k and \hat{a}_k^\dagger , then using (B.69) we find that we should impose

$$u_k u_k^{*\prime} - u_k^* u_k' = i, \quad (10.81)$$

i.e. that the modes u_k are not normalized arbitrarily.

It is interesting to notice that for the modes satisfying the linear second-order equation (10.79), the (10.81) simply expresses the value of the Wronskian of the solutions of this equation, a value that is conserved in time by the properties of this Wronskian. In other words, we see that a special normalization should be imposed on the modes, and that this normalization is preserved in time.

10.2.2.4 Choice of integration constants

Equation (10.79), or its equivalent in other situations of interest, is a linear second-order differential equation. Its general solution therefore has two a priori arbitrary integration constants, related by the Wronskian constraint (10.81).

A second relation can be found if there is some (space or time-like) asymptotic regions for which the positive frequency modes are well defined. For instance, if the space-time is, up to a good approximation, of Minkowski type; we will see an exact illustration of this type later. In this case, the scale factor is constant and the curvature vanishes, so that (10.79) reduces to the usual case $u'' + \omega^2 u = 0$, whose solutions are $\exp(\pm i\omega\eta)$. The solution with the + sign (respectively, -) corresponds to the negative (resp. positive) frequency. We thus require that the annihilation operator \hat{a}_k be associated with positive frequency and that leads us to keep only the solution with the - sign. Taking now into account the Wronskian normalization, we find that we should take

$$u_k = \frac{e^{-i\omega\eta}}{\sqrt{2\omega}}, \quad (10.82)$$

which is the usual Minkowski space-time relation and is useful in cosmology to fix the initial conditions (see Chapter 8).

10.2.2.5 Bunch–Davies choice for de Sitter

We now consider the particular case of a de Sitter space-time, for which the scale factor is given by

$$a(\eta) = \frac{1}{-H\eta}, \quad -\infty < \eta < 0. \quad (10.83)$$

For the solution (10.83), the scalar curvature is expressed simply by $R = -12H^2$, so that the solution of (10.79) is

$$u_k(\eta) = \frac{\sqrt{-\pi\eta}}{2} \left[A_1(k) e^{i\frac{\pi}{2}(\nu+\frac{1}{2})} H_\nu^{(1)}(-k\eta) + A_2(k) e^{-i\frac{\pi}{2}(\nu+\frac{1}{2})} H_\nu^{(2)}(-k\eta) \right], \quad (10.84)$$

where $A_1(k)$ and $A_2(k)$ are arbitrary functions of the wavenumber k , $H_\nu^{(1),(2)}(k|\eta|)$ the Hankel functions¹ of index $\nu = \frac{9}{4} - m^2/H^2 - 12\xi$.

With this normalization, the relation for the Wronskian (10.81) simply becomes

$$|A_1(k)|^2 - |A_2(k)|^2 = 1,$$

and we should now impose the positive frequency criteria to fix the choice in an unambiguous way. To do so, we are interested in the asymptotic expansion of the Hankel functions, and we find that with the phase choice (10.84), we have

$$u_k(\eta) \sim \frac{1}{\sqrt{2k}} (A_1 e^{-ik\eta} + A_2 e^{ik\eta}),$$

which indicates that we should take $A_2 = 0$ to keep only the positive frequency, and due to the normalization, this imposes $A_1 = 1$ up to a phase.

The relevant solution in de Sitter space is thus

$$u_k(\eta) = \frac{\sqrt{-\pi\eta}}{2} e^{i\frac{\pi}{2}(\nu+\frac{1}{2})} H_\nu^{(1)}(-k\eta),$$

(10.85)

leading to the so-called *Bunch–Davies vacuum state* [23].

10.2.2.6 Bogoliubov coefficients

Considering the general case (10.78) of a field decomposition of the form

$$\hat{\phi} = \sum_i \left[\hat{a}_i u_i(x) + \hat{a}_i^\dagger u_i^*(x) \right], \quad (10.86)$$

where the sum is over all the degrees of freedom of the decomposition: a discrete sum over the discrete indices, an integral over the continuous indices. A vacuum state $|0\rangle_u$, is associated to this representation, annihilated by all the operators \hat{a}_i : $\forall i$, $\hat{a}_i |0\rangle_u = 0$. Using this vacuum state as a basis, we can then construct the entire Fock space corresponding to the states that contain particles.

Unlike in the flat case, there is, in general, no time-like Killing vector with positive-frequency eigenmodes u_i . The decomposition (10.86) is not special like its flat space-time counterpart, and in particular we can choose another basis of modes

$$\hat{\phi} = \sum_j \left[\hat{b}_j v_j(x) + \hat{b}_j^\dagger v_j^*(x) \right]. \quad (10.87)$$

In the absence of any reason to the contrary, this second description is as valid as (10.86). In particular, these new modes define a new vacuum state $|0\rangle_v$, annihilated by the operators \hat{b}_j , and a different Fock space from the first one.

¹We should note here that since the conformal time is negative, we have chosen the Hankel functions with positive argument, hence the sign $-k\eta$, so as to avoid any cutoff problem along the negative real axis for the Hankel functions.

$$u_k(\eta) = \frac{\sqrt{-\pi\eta}}{2} \left[A_1(k) e^{i\frac{\pi}{2}(\nu - \frac{1}{2})} H_\nu^{(1)}(-k\eta) + A_2(k) e^{-i\frac{\pi}{2}(\nu + \frac{1}{2})} H_\nu^{(2)}(-k\eta) \right], \quad (10.84)$$

where $A_1(k)$ and $A_2(k)$ are arbitrary functions of the wavenumber k , $H_\nu^{(1),(2)}(k|\eta|)$ the Hankel functions¹ of index $\nu = \frac{9}{4} - m^2/H^2 - 12\xi$.

With this normalization, the relation for the Wronskian (10.81) simply becomes

$$|A_1(k)|^2 - |A_2(k)|^2 = 1,$$

and we should now impose the positive frequency criteria to fix the choice in an unambiguous way. To do so, we are interested in the asymptotic expansion of the Hankel functions, and we find that with the phase choice (10.84), we have

$$u_k(\eta) \sim \frac{1}{\sqrt{2k}} (A_1 e^{-ik\eta} + A_2 e^{ik\eta}),$$

which indicates that we should take $A_2 = 0$ to keep only the positive frequency, and due to the normalization, this imposes $A_1 = 1$ up to a phase.

The relevant solution in de Sitter space is thus

$$u_k(\eta) = \frac{\sqrt{-\pi\eta}}{2} e^{i\frac{\pi}{2}(\nu + \frac{1}{2})} H_\nu^{(1)}(-k\eta),$$

(10.85)

leading to the so-called *Bunch–Davies vacuum state* [23].

10.2.2.6 Bogoliubov coefficients

Considering the general case (10.78) of a field decomposition of the form

$$\hat{\phi} = \sum_i \left[\hat{a}_i u_i(x) + \hat{a}_i^\dagger u_i^*(x) \right], \quad (10.86)$$

where the sum is over all the degrees of freedom of the decomposition: a discrete sum over the discrete indices, an integral over the continuous indices. A vacuum state $|0\rangle_u$, is associated to this representation, annihilated by all the operators \hat{a}_i : $\forall i, \hat{a}_i |0\rangle_u = 0$. Using this vacuum state as a basis, we can then construct the entire Fock space corresponding to the states that contain particles.

Unlike in the flat case, there is, in general, no time-like Killing vector with positive-frequency eigenmodes u_i . The decomposition (10.86) is not special like its flat space-time counterpart, and in particular we can choose another basis of modes

$$\hat{\phi} = \sum_j \left[\hat{b}_j v_j(x) + \hat{b}_j^\dagger v_j^*(x) \right]. \quad (10.87)$$

In the absence of any reason to the contrary, this second description is as valid as (10.86). In particular, these new modes define a new vacuum state $|0\rangle_v$, annihilated by the operators \hat{b}_j , and a different Fock space from the first one.

¹We should note here that since the conformal time is negative, we have chosen the Hankel functions with positive argument, hence the sign $-k\eta$, so as to avoid any cutoff problem along the negative real axis for the Hankel functions.

Both sets of modes u_i and v_j describe the same system, and both of them must be complete. As a result, it is possible to express each one of these modes as a function of those from the other basis. We have the Bogoliubov relations [24]

$$v_j(x) = \sum_i [\alpha_{ji} u_i(x) + \beta_{ji} u_i^*(x)], \quad u_i(x) = \sum_j [\alpha_{ji}^* v_j(x) - \beta_{ji} v_j^*(x)], \quad (10.88)$$

in which the Bogoliubov coefficients α_{ij} and β_{ij} are given by

$$\alpha_{ij} = \langle v_i, u_j \rangle \quad \text{and} \quad \beta_{ij} = -\langle v_i, u_j^* \rangle, \quad (10.89)$$

with the scalar product of (10.77), and using the property that $\langle g, f \rangle = \langle f, g \rangle^*$.

Writing explicitly the equality between the decompositions (10.86) and (10.87), and identifying them term by term, we find

$$\hat{b}_j = \sum_i (\alpha_{ji}^* \hat{a}_i - \beta_{ji} \hat{a}_i^\dagger) \quad \text{and} \quad \hat{a}_i = \sum_j (\alpha_{ji} \hat{b}_j + \beta_{ji}^* \hat{b}_j^\dagger), \quad (10.90)$$

which relates the creation and annihilation operators from one basis to another.

Moreover, the Bogoliubov coefficients also have the following properties

$$\sum_k (\alpha_{ik} \alpha_{jk}^* - \beta_{ik} \beta_{jk}^*) = \delta_{ij} \quad \text{and} \quad \sum_k \alpha_{ik} \beta_{jk} = \sum_k \beta_{ik} \alpha_{jk}, \quad (10.91)$$

these relations are obtained by performing the initial transformations (10.88) twice.

Both Fock spaces defined by the operators \hat{a} and \hat{b} do not lead to the same description in terms of particles. Indeed, the vacuum state $|0\rangle_a$ is not annihilated by \hat{b} , since

$$\hat{b}_j |0\rangle_a = - \sum_i \beta_{ji} \hat{a}_i^\dagger |0\rangle_a \neq 0,$$

and similarly for $\hat{a}|0\rangle_b$. As a result, the vacuum state for the expansion \hat{a} is not the vacuum for the expansion \hat{b} as soon as the Bogoliubov coefficients $\beta_{ji} \neq 0$. Actually, denoting by $\hat{\mathcal{N}}_j^{(b)} = \hat{b}_j^\dagger \hat{b}_j$ (without summing over the repeated index) the number of particles associated with the \hat{b} operator, then its value in the vacuum state of \hat{a} is

$${}_a\langle 0 | \hat{\mathcal{N}}_j^{(b)} | 0 \rangle_a = \sum_k |\beta_{jk}|^2. \quad (10.92)$$

So, the vacuum of the u modes contain particles of v modes. We will now study an exact example illustrating this mechanism of particle creation by a gravitational field.

10.2.2.7 Choice of the initial state

We often find in the literature that the choice (10.82) for Minkowski or (10.85) for de Sitter (or their equivalent, if it exists, in the space we are interested in) amounts to saying that the initial state of the system is the vacuum. In itself, this comment has no content since the expansion of the field into creation and annihilation operators of the field excitation quanta indicates nothing for the state in which the system will

be found. In practice, to be meaningful [25], one should also specify the state of the system, and we assume that it is $|0\rangle_u$, i.e. the state annihilated by the u modes.

Taking the expansion (10.86), we find that the expectation value $\langle\hat{\phi}^2\rangle$ in the vacuum of the u modes is

$$\langle\hat{\phi}^2\rangle \equiv {}_u\langle 0|\hat{\phi}^\dagger\hat{\phi}|0\rangle_u = \sum_i |u_i|^2. \quad (10.93)$$

As a result, different choices of integration constants for the u_i modes lead to different results for the vacuum expectation value (10.93), which can also be interpreted as different choices of vacuum for the theory. Once this choice is made, we can still decide to impose as an initial condition that the system is in the corresponding vacuum state. What is often written actually then amounts to deciding a priori that the system is in the vacuum, and then fixing this vacuum by the choice of mode integration constants.

10.2.3 A complete example

We discuss here an example initially studied in Ref. [26] for which we reproduce the procedure of Ref. [20].

10.2.3.1 Two-dimensional metric

We consider the case of a minimally coupled scalar field with mass m in a two-dimensional space-time. The metric of Friedmann–Lemaître type is

$$ds^2 = -dt^2 + a^2(t)dx^2 = a^2(\eta) (-d\eta^2 + dx^2).$$

We choose for the scale factor a solution (the scalar field is quantized in an exterior gravitational field, and we assume that the scale factor is a solution of the classical Einstein equations), given by

$$a(\eta) = A + B\tanh\left(\frac{\eta}{\eta_0}\right), \quad (10.94)$$

where η_0 reflects the typical length during which the scale factor evolves. For $\eta \gg \eta_0$, we have both limits $\lim_{\eta \rightarrow \pm\infty} a = A \pm B$. Asymptotically, the metrics in regions $\eta \gg \eta_0$ tend towards two-dimensional Minkowski metrics.

For $\xi = 0$, (10.74) becomes

$$\frac{d^2 u_k}{d\eta^2} + [k^2 + a^2(\eta)m^2] u_k = 0, \quad (10.95)$$

where we have used the expansion

$$\hat{\phi}(x, \eta) = \int \frac{dk}{\sqrt{2\pi}} \left[\hat{a}_k Q_k(x) u_k(\eta) + \hat{a}_k^\dagger Q_k^*(x) u_k^*(\eta) \right], \quad (10.96)$$

with $Q_k(x) = e^{ikx}$, the one-dimensional equivalent of (10.80).

10.2.3.2 Incoming and outgoing modes

For the scale factor (10.94), it turns out that the general solution of (10.95) is known [26]; like any solution of a linear second-order differential equation, it depends on two integration constants, which are chosen so as to have modes of positive frequency as limits for $\eta \rightarrow -\infty$ or $\eta \rightarrow +\infty$. Once this choice is made, the normalization (10.81) gives another relation between the constants, so that the solution is completely determined.

Proceeding in this manner for $\eta \rightarrow -\infty$, we find

$$u_k^{\text{in}}(\eta) = \frac{e^{-i\omega_+ \eta - i\omega_- \eta_0 \ln[2\cosh(\eta/\eta_0)]}}{\sqrt{2\omega_{\text{in}}}} {}_2F_1(1 + i\omega_- \eta_0, i\omega_- \eta_0, 1 - i\omega_{\text{in}} \eta_0; z_+), \quad (10.97)$$

where ${}_2F_1$ is a hypergeometric function (see Refs. [2, 3] of Appendix B), and for $\eta \rightarrow +\infty$, we obtain

$$u_k^{\text{out}}(\eta) = \frac{e^{-i\omega_+ \eta - i\omega_- \eta_0 \ln[2\cosh(\eta/\eta_0)]}}{\sqrt{2\omega_{\text{in}}}} {}_2F_1(1 + i\omega_- \eta_0, i\omega_- \eta_0, 1 + i\omega_{\text{out}} \eta_0; z_-). \quad (10.98)$$

The constants that appear in these relations are

$$\omega_{\text{in}}^2 = k^2 + m^2(A - B), \quad \omega_{\text{out}}^2 = k^2 + m^2(A + B), \quad \text{and} \quad \omega_{\pm} = \frac{1}{2}(\omega_{\text{out}} \pm \omega_{\text{in}}),$$

and the variable is defined by

$$z_{\pm} = \frac{1}{2} \left[1 \pm \tanh \left(\frac{\eta}{\eta_0} \right) \right].$$

As required, both modes have the limit

$$\lim_{\eta \rightarrow -\infty} u_k^{\text{in}}(\eta) = \frac{e^{-i\omega_{\text{in}} \eta}}{\sqrt{2\omega_{\text{in}}}}, \quad \lim_{\eta \rightarrow +\infty} u_k^{\text{out}}(\eta) = \frac{e^{-i\omega_{\text{out}} \eta}}{\sqrt{2\omega_{\text{out}}}},$$

satisfying the Wronskian constraint (10.81).

From the properties of the hypergeometric functions, we can deduce the relations between the incoming modes, ‘in’, and outgoing modes, ‘out’, in the form

$$\begin{aligned} u_k^{\text{in}}(\eta) &= \int dp [\alpha_{kp} u_p^{\text{out}}(\eta) + \beta_{kp} u_p^{\text{out}*}(\eta)] \\ &= \int dp [\alpha_k \delta(k - p) u_p^{\text{out}}(\eta) + \beta_k \delta(k + p) u_p^{\text{out}*}(\eta)], \end{aligned} \quad (10.99)$$

with

$$\alpha_k = \sqrt{\frac{\omega_{\text{out}}}{\omega_{\text{in}}}} \frac{\Gamma(1 - i\omega_{\text{in}} \eta_0) \Gamma(-i\omega_{\text{out}} \eta_0)}{\Gamma(-i\omega_+ \eta_0) \Gamma(1 - i\omega_+ \eta_0)}, \quad \beta_k = \sqrt{\frac{\omega_{\text{out}}}{\omega_{\text{in}}}} \frac{\Gamma(1 - i\omega_{\text{in}} \eta_0) \Gamma(i\omega_{\text{out}} \eta_0)}{\Gamma(i\omega_- \eta_0) \Gamma(1 + i\omega_+ \eta_0)},$$

giving the Bogoliubov coefficients for this system in the form

$$\alpha_{kp} = \alpha \delta(k - p) \quad \text{and} \quad \beta_{kp} = \beta_k \delta(k + p).$$

10.2.3.3 Particles production

Taking the modulus of the Bogoliubov coefficients, and using the properties of the Γ functions, we obtain

$$|\alpha_k|^2 = \frac{\operatorname{sh}^2(\pi\omega_+\eta_0)}{\operatorname{sh}(\pi\omega_{\text{in}}\eta_0)\operatorname{sh}(\pi\omega_{\text{out}}\eta_0)} \quad \text{and} \quad |\beta_k|^2 = \frac{\operatorname{sh}^2(\pi\omega_-\eta_0)}{\operatorname{sh}(\pi\omega_{\text{in}}\eta_0)\operatorname{sh}(\pi\omega_{\text{out}}\eta_0)},$$

from which we can use the properties (10.91) to check that

$$|\alpha_k|^2 - |\beta_k|^2 = 1.$$

Assuming now that the field ϕ is in the vacuum state at $\eta \rightarrow \infty$, i.e. $|0\rangle_{\text{in}}$. In the Heisenberg representation, for which the states do not evolve, this state still describes the system at all times. Originally, since this vacuum state is defined with respect to the u_k^{in} modes, there are no particles, and this is true for any inertial observer since in this limit the space-time is Minkowski. In the region corresponding to the outgoing modes, the situation is similar. The space-time is again Minkowski, but with a different distance normalization, and the notion of particles is thus ensured to have a clear meaning. Since the state $|0\rangle_{\text{in}}$ is not the vacuum of the u_k^{out} modes, all inertial observers will see a number of particles equal to $|\beta_k|^2$. The expansion of the Universe, in this simple case, has produced scalar field particles from the coupling with gravity.

It is via this mechanism that perturbations of a scalar field are produced, giving rise, in the post-inflationary primordial Universe, to the classical fluctuations required for the formation of the large-scale structures (see Chapter 8).

10.3 Supersymmetry and supergravity

Before we formulate particle physics in a curved space-time along the lines described previously, particle physics must first be described in a Minkowski space-time, which is always possible locally from the principles of general relativity. The description of particles amounts to studying microscopically the different symmetries that exist between the physical states, and to see these effects it is almost always possible to neglect gravity, except in the cases for which field theory in curved space is not valid: this occurs when the characteristic energy of the systems we consider is of the order of M_p . As well as gauge symmetries, which can be arbitrarily extended to the grand unified theories as seen in Chapter 9, there is an additional symmetry, known as supersymmetry that, whilst being complementary to the others, is not itself based on a Lie group. According to this symmetry, fermions and bosons are only different states of the same mathematical object known as a supermultiplet.

10.3.1 Technical generalities

The supersymmetry principle, often denoted by SUSY, assumes the existence of a fermionic operator Q that exchanges the fermion and boson states:

$$Q|\text{boson}\rangle = |\text{fermion}\rangle \quad \text{and} \quad Q|\text{fermion}\rangle = |\text{boson}\rangle,$$

and postulates that these operators are the generators of a symmetry. By doing so, a new particle should be introduced, which can be motivated both from experimental and theoretical considerations.

10.3.1.1 Motivations

There are many independent pieces of evidence that point to the existence of supersymmetry. Furthermore, these reasons indicate that the characteristic energy scale should be of the order of a TeV. First, grand unification: as discussed earlier, the coupling constants of electromagnetic, weak and strong interactions seem to join at high energy, but not exactly, as shown in Fig. 9.2. This unification can be greatly improved by assuming supersymmetry, as can be observed in Fig. 10.5, provided its effects are only noticeable at energies above a typical scale of the order of a TeV.

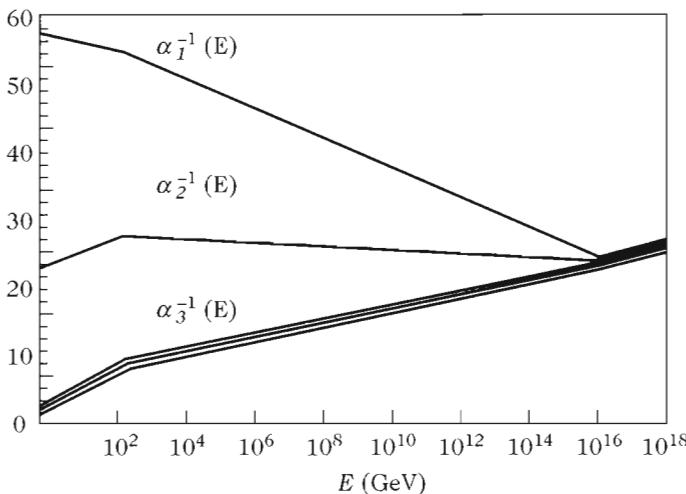


Fig. 10.5 Unification of the coupling constants with supersymmetry. If supersymmetry is present, the number of particles contributing to interactions is greater, changing the slope of the variation of the coupling constants with the energy beyond the supersymmetry-breaking scale. This provides us with a way to modify the slopes so that the constants intersect exactly at the same energy. Furthermore, this allows us to predict the energy scale at which supersymmetry is broken, in this case around 1 TeV.

Moreover, supersymmetry gives a physical solution to an apparently technical problem: the hierarchy problem, i.e. the issue of why the masses of scalar particles, such as the Higgs boson of the standard model, are not of the order of grand unification. Considering radiative corrections δm to the mass m of a scalar particle, we notice that they depend on the energy scale beyond which the theory is no longer valid, for example, E_{GUT} . We then find that $m^2 = m_0^2 + \delta m^2 \sim m_0^2 - E_{\text{GUT}}^2$, so that to have, for instance, $m \sim 100$ GeV, we should have $(m_0^2 - E_{\text{GUT}}^2)/E_{\text{GUT}}^2 \simeq 10^{-34}$, which is quite a considerable fine tuning of parameters that originally had nothing to do with each other.

As well as the usual particles, supersymmetry implies the existence of supersymmetric partners of different spin; these are known as superpartners, *sparticles* or *spar-*

ticles on the radiative corrections is to compensate the effect of the initial particles, so as to exactly cancel these corrections. This resolves the hierarchy problem. Moreover, if supersymmetry is broken, even though the compensation is no longer exact, it remains very good, and the problem is far less severe: whichever way, if the hierarchy problem is solved at a given order, then this solution remains valid to all orders in perturbations. In order for this to be valid, supersymmetry should be able to act on energy scales not very different from that involved in the standard model, i.e. of the order of a hundred GeV.

Finally, the last argument arises from cosmology: for realistic models, there must exist an absolutely stable particle called the lightest supersymmetric particle (LSP). In order for this particle to contribute significantly to dark matter, its mass should be of the order of a hundred of GeV, hence once more the right energy scale.

To conclude, most realistic theories assume the existence of a supersymmetry, which has to be broken for energy scales below the TeV.

10.3.1.2 *Coleman–Mandula and Haag–Lopuszański–Sohnius theorems*

The first problem that arises when one wants to construct a supersymmetric theory stems from the Coleman–Mandula theorem of 1967 [27]. According to it, any relativistic field theory in four dimensions that is invariant under the transformations of a semi-simple Lie group and contains the Poincaré group as a subgroup is a trivial theory: the scattering matrix is so constrained as to respect the conditions of the theorem that it can only be equal to the identity. In other words, there are no interactions, and so the theory, being trivial, is of no physical interest.

This theorem can also be seen as determining the largest symmetry a non-trivial theory can have. It is the product of the Poincaré group with an internal symmetry group. In this formulation we see that the theorem implies that it is not possible for irreducible supermultiplets of the symmetry group to contain particles of different spins. Indeed, in order for such a multiplet to contain both bosons and fermions, there must exist a generator of the invariance group capable of transforming one into another. Since the spin is related to rotational invariance and since the elements of the Poincaré group leave it invariant, the invariance required for the theory should be a semi-simple extension of the Poincaré group.

This result is actually based on the assumption that the new generators introduced, i.e. those of the supersymmetry, form a Lie algebra. In 1971, GolFang and Likhtman [28] noticed that in reality, states with different spins could be related via a symmetry group whose generators form a graded Lie algebra, i.e. they satisfy anticommutation as well as commutation relations. Finally, in 1975, Haag, Lopuszański and Sohnius [29] demonstrated that these operators could not have a spin greater than $\frac{1}{2}$ for the theory to be non-trivial. The supersymmetry operators must therefore be spinors.

10.3.1.3 *Weyl, Dirac and Majorana spinors*

In the chiral representation (2.65) used so far, a Dirac spinor is decomposed in the form of two Weyl spinors as

$$\psi = \begin{pmatrix} \xi_\alpha \\ \bar{\chi}^{\dot{\alpha}} \end{pmatrix} \quad (10.100)$$

where the undotted indices $\alpha = 1, 2$ give the components of a left-handed spinor and the dotted indices $\dot{\alpha} = 1, 2$ indicate the right-handed components. These differences in the indices distinguished the two kinds of spinors that transform differently under Lorentz transformations. Moreover, it turns out [30] that the Weyl spinors $\sigma^2 \xi^*$ and $\sigma^2 \bar{\chi}^*$ transform, respectively, as $\bar{\chi}$ and ξ , so that it is possible to transform a left-handed bi-spinor into a right-handed one and vice versa thanks to the Pauli matrix σ^2 . Defining $\bar{\xi}_{\dot{\alpha}} \equiv (\xi_\alpha)^*$ and $\chi^\alpha = (\bar{\chi}^{\dot{\alpha}})^*$, we can use the matrices $\epsilon_{\alpha\beta} = (i\sigma^2)_{\alpha\beta} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ and $\epsilon^{\dot{\alpha}\dot{\beta}} = (-i\sigma^2)^{\dot{\alpha}\dot{\beta}} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ to pass from a right spinor to a left spinor and vice versa: $\chi_\alpha = \epsilon_{\alpha\beta} \chi^\beta$ and $\bar{\xi}^{\dot{\alpha}} = \epsilon^{\dot{\alpha}\dot{\beta}} \bar{\xi}_{\dot{\beta}}$. In the same manner, we can define complementary operators to raise and lower the indices by $\epsilon_{\dot{\alpha}\dot{\beta}} = \epsilon_{\alpha\beta}$ and $\epsilon^{\alpha\beta} = \epsilon^{\dot{\alpha}\dot{\beta}}$, giving $\chi^\alpha = \epsilon^{\alpha\beta} \chi_\beta$ and $\bar{\xi}_{\dot{\alpha}} = \epsilon_{\dot{\alpha}\dot{\beta}} \bar{\xi}^{\dot{\beta}}$. Finally, with these notations the conjugate $\bar{\psi}$ of a Dirac spinor is $\bar{\psi} = (\chi^\alpha, \bar{\xi}_{\dot{\alpha}})$.

The charge conjugation operator \hat{C} for the fermions is [31] $\hat{C} = -i\gamma^0\gamma^2$ so that the charge conjugate fermion ψ^c of ψ is

$$\psi^c = \hat{C} \bar{\psi}^T, \quad \hat{C} = -i\gamma^0\gamma^2 = \begin{pmatrix} -i\sigma^2 & 0 \\ 0 & i\sigma^2 \end{pmatrix}, \quad (10.101)$$

which is used to define a Majorana fermion ψ_M by imposing $\psi_M^c = \psi_M$. One can check explicitly that if ψ satisfies the Dirac equation for a charged particle of charge q , then ψ^c satisfies the same equation as long as we replace q by $-q$. A Majorana spinor can be simply constructed from a Weyl spinor in the form $\psi_M = \begin{pmatrix} \psi_\alpha \\ \bar{\psi}^{\dot{\alpha}} \end{pmatrix}$.

With these notations, the Dirac Lagrangian of a field of mass m (2.63) becomes

$$\mathcal{L}_{\text{Dirac}} = -i\chi^\alpha (\sigma^\mu)_{\alpha\dot{\beta}} \partial_\mu \bar{\chi}^{\dot{\beta}} - i\bar{\xi}_{\dot{\alpha}} (\bar{\sigma}^\mu)^{\dot{\alpha}\dot{\beta}} \partial_\mu \xi_\beta - m(\chi^\alpha \xi_\alpha + \bar{\xi}_{\dot{\alpha}} \bar{\chi}^{\dot{\alpha}}),$$

where the indices of the matrices σ^μ and $\bar{\sigma}^\mu$ are adjusted so that the Lagrangian is a scalar. To go further in the computations, we should take into account the fact that the fermions, once quantized, satisfy anticommutation relations [see (2.114) and (2.115)]. For the classical theory,² this translates into the fact that we require that all the spinor components anticommute rather than commute. We call them *Grassmann variables*.

The first term in the Dirac Lagrangian can also be written in the form

$$-i\chi^\alpha (\sigma^\mu)_{\alpha\dot{\beta}} \partial_\mu \bar{\chi}^{\dot{\beta}} = \partial_\mu \left[-i\chi^\alpha (\sigma^\mu)_{\alpha\dot{\beta}} \bar{\chi}^{\dot{\beta}} \right] + i\partial_\mu \chi^\alpha (\sigma^\mu)_{\alpha\dot{\beta}} \bar{\chi}^{\dot{\beta}}.$$

²A classical theory for a fermionic field is not strictly speaking well defined, since a fermion is quantized by nature. We use this designation to describe the theory on which the correspondence principle is applied to obtain the quantum theory (see Chapter 2). This formulation is useful in particular in the functional formulation, i.e. in the path integral approach, in field theory.

Since the total derivative term has no influence at the level of the dynamical equations, we find that by taking into account the relation $\sigma^2 \bar{\sigma}^\mu \sigma^2 = (\bar{\sigma}^\mu)^T$ and the fact that $\partial_\mu \chi_\beta \bar{\chi}_\dot{\alpha} = -\bar{\chi}_{\dot{\alpha}} \partial_\mu \chi_\beta$, we obtain

$$\mathcal{L}_{\text{Dirac}} = -i(\bar{\xi} \bar{\sigma}^\mu \xi + \bar{\chi} \bar{\sigma}^\mu \chi) + m(\chi \xi + \bar{\chi} \bar{\xi}), \quad (10.102)$$

where the ‘scalar product’ for the fermions accounts for the anticommutativity: $\chi \xi \equiv \chi^\alpha \xi_\alpha = \xi^\alpha \chi_\alpha = \xi \chi$ and $\bar{\chi} \bar{\xi} \equiv \bar{\xi}_{\dot{\alpha}} \bar{\chi}^{\dot{\alpha}} = \bar{\chi}_{\dot{\alpha}} \bar{\xi}^{\dot{\alpha}} = \bar{\chi} \bar{\xi}$. Equation (10.102) shows that the action that describes the Dirac spinors can always be written by expanding the latter in the form of a collection of left Weyl spinors.

10.3.1.4 Graded algebra

The simplest possibility for the supersymmetry generators allowing for the transformation of fermions states into boson states is the one for which we choose the operators Q_α^A to be Weyl bi-spinors, with $A = 1, \dots, N$, where N is the number of supersymmetries. We also assume an internal symmetry between the different generators Q_α^A , which are thus placed into a representation fixed by this invariance. With this choice, we can then show [31] that the supersymmetry algebra is written in the form

$$[P^\mu, Q_\alpha^A] = 0, \quad (10.103)$$

indicating that the supersymmetry transformations are not affected by translations, which have the generator $P^\mu = i\partial_\mu$, and

$$[J^{\mu\nu}, Q_\alpha^A] = -i(\sigma^{\mu\nu})_\alpha^\beta Q_\beta^A, \quad \text{and} \quad \left[J^{\mu\nu}, \bar{Q}_A^{\dot{\alpha}} \right] = -i(\bar{\sigma}^{\mu\nu})_{\dot{\beta}}^{\dot{\alpha}} \bar{Q}_A^{\dot{\beta}}, \quad (10.104)$$

where $\sigma^{\mu\nu} = \frac{1}{4}(\sigma^\mu \bar{\sigma}^\nu - \sigma^\nu \bar{\sigma}^\mu)$ and $\bar{\sigma}^{\mu\nu} = \frac{1}{4}(\bar{\sigma}^\mu \sigma^\nu - \bar{\sigma}^\nu \sigma^\mu)$ and the operators $J_{\mu\nu}$ are the generators of rotations and Lorentz transformations whose properties are discussed in Chapter 1 [see (1.16)]. Moreover, we have

$$\left\{ Q_\alpha^A, \bar{Q}_{\dot{\beta}}^B \right\} = 2\delta_B^A (\sigma^\mu)_{\alpha\dot{\beta}} P_\mu, \quad (10.105)$$

and finally

$$\left\{ Q_\alpha^A, Q_\beta^B \right\} = \epsilon_{\alpha\beta} Z^{AB}, \quad (10.106)$$

where the ‘central charges’ Z^{AB} satisfy

$$Z^{AB} = -Z^{BA} = \sum_a (q^{AB})_a \lambda^a,$$

the λ^a being the generators of the internal symmetries of the supersymmetry generators respecting the algebra (2.119). These central charges commute with all the supersymmetry generators: they belong to an Abelian subalgebra invariant under the internal symmetry group.

The algebra formed by (2.119) together with the commutation relations (1.15), (1.16) of the Poincaré group, and supplemented with (10.103) to (10.106) include both commutation and anticommutation relations. So these relation do not a Lie

algebra, which explains why the Coleman–Mandula theorem does not apply. This type of algebra is called a *graded algebra*.

In many practical cases, in particular in cosmology, we consider the case $N = 1$, which then has no central charges. If $N \neq 1$, we are dealing with the so-called extended supersymmetry. From now on, we restrict ourselves to the case $N = 1$ SUSY.

10.3.1.5 Positivity of the energy

An important consequence of supersymmetry relates to the energy of the state of the system. Using (10.105) and taking the trace over the spin indices, then, since the trace of the Pauli matrices vanishes, we find

$$Q_1 \bar{Q}_1 + \bar{Q}_1 Q_1 + Q_2 \bar{Q}_2 + \bar{Q}_2 Q_2 = \sum_{\alpha=1}^2 [Q_\alpha (Q_\alpha)^* + (Q_\alpha)^* Q_\alpha] = 4P_0.$$

Now, if we consider an arbitrary physical state $|\Psi\rangle$, its Hamiltonian $\hat{H} = P_0$ necessarily satisfies

$$\langle \Psi | \hat{H} | \Psi \rangle = \frac{1}{4} \sum_{\alpha} [| (Q_\alpha)^* |\Psi\rangle |^2 + | |Q_\alpha| \Psi \rangle |^2] \geq 0, \quad (10.107)$$

which is a sum of positive-definite terms. This implies that the energy of the state $|\Psi\rangle$ is manifestly positive. Therefore, we find that any state of a supersymmetric theory has positive energy, and we see that the minimum of the energy, i.e. the vacuum $|0\rangle$ is obtained for $Q_\alpha|0\rangle = 0 = Q_\alpha^*|0\rangle$. The supersymmetric vacuum thus satisfies $H|0\rangle = 0$.

This property is useful in many quantum systems, even in non-relativistic ones [32], as well as in cosmology.

10.3.1.6 Supersymmetry multiplets

As for any symmetry, the irreducible representations of supersymmetry are constructed on the basis of the eigenvalues of the Casimir operators, i.e. the operators that commute with all the generators of the invariance group. This is, for instance, the case for the operator $P^2 = P_\mu P^\mu$ in the Poincaré group, whose eigenvalue, here the squared mass, is the same for all the particles of the same representation.

In supersymmetry, (10.103) shows that P^2 also commutes with the generators Q_α , so it is still a Casimir operator, and we can thus still classify the representations according to its eigenvalues, that is the mass: the particles and their superpartners must have the same mass.

Another interesting operator is the Pauli–Lubanski spin vector W^μ defined by

$$W^\mu \equiv \frac{1}{2} \varepsilon^{\mu\nu\alpha\beta} P_\nu J_{\alpha\beta},$$

the square³ of which is a Casimir operator for the Poincaré group: $W^2 = m^2 \mathbf{J}^2$, with $m^2 = -P_\mu P^\mu$ and \mathbf{J} the angular momentum, with eigenvalues $\mathbf{J}^2 = j(j+1)$: the

³The square of W_μ is computed using the relation [see Ref. (1.32) of Chapter 1] as well as using the explicit relations [9] of Chapter 2 $J_{ij} = \varepsilon_{ij}^{k} J_k$ and $J_{0i} = K_i$, where \mathbf{J} is the rotation generator (angular momentum) and \mathbf{K} that of the Lorentz transformations.

irreducible representations in non-supersymmetric physics are classified according to the particle spins. As soon as the fermionic generators are taken into account, the direct computation shows that $[W^2, Q_\alpha] \neq 0$, which implies that a supersymmetry multiplet, i.e. a supermultiplet, can contain massive particles of different spins.

Actually, each supermultiplet must contain as many bosonic as fermionic degrees of freedom. This can be seen in the following way: the operators Q_α and $\bar{Q}_\dot{\alpha}$ change a bosonic state into a fermionic state and vice versa. Thus, the anticommutator $\{Q_\alpha, \bar{Q}_\dot{\alpha}\}$, is a bosonic operator, which is proportional to the momentum P_μ from the algebra relation (10.105). Denoting by N_F the fermionic number operator, which gives 1 on a fermion state and 0 on a boson state, we therefore have $(-1)^{N_F} Q_\alpha = -Q_\alpha (-1)^{N_F}$. On a supermultiplet of finite dimension, we find that $\text{Tr} [(-1)^{N_F} \{Q_\alpha, \bar{Q}_\beta\}] = 0$, which implies from (10.105) that $\text{Tr} [(-1)^{N_F}] = 0$. Since $(-1)^{N_F} = -1$ on a fermion state and $(-1)^{N_F} = +1$ on a boson state, this implies that there must be as many fermions as bosons in a supermultiplet.

10.3.1.7 Infinitesimal transformations

We can realize the supersymmetry algebra using an infinitesimal transformation, having a ‘small’ fermionic parameter ζ^α , which must be a Grassmann variable. The limit of a large number of such infinitesimal transformations, as in the case of (2.117), leads to a finite transformation. Such a transformation is generated by

$$T_\zeta = 1 - i \left(\zeta^\alpha Q_\alpha + \bar{\zeta}_{\dot{\alpha}} \bar{Q}^{\dot{\alpha}} \right), \quad (10.108)$$

and we find [31] that the transformation of an arbitrary fermionic or bosonic field A is given by the commutator

$$\delta_\zeta A(x) = [i (\zeta Q + \bar{\zeta} \bar{Q}), A(x)]. \quad (10.109)$$

In order for the algebra (10.103)–(10.106) to be satisfied, the infinitesimal transformations must then satisfy closure relations of the form $[\zeta Q, \eta Q] = [\zeta \bar{Q}, \eta \bar{Q}] = 0$, and especially

$$[\zeta Q, \bar{\eta} \bar{Q}] = 2\zeta \sigma^\mu \bar{\eta} P_\mu, \quad (10.110)$$

indicating how two successive transformations must act in order to respect the supersymmetry.

10.3.2 Wess–Zumino model

The simplest supersymmetric model we can conceive is based on a left Weyl fermion, containing two degrees of freedom, which can thus be balanced in the bosonic sector by a complex scalar field. This is the Wess–Zumino model [33].

10.3.2.1 Kinetic terms

We consider the following kinetic terms

$$\mathcal{L}_{wz}^{\text{kin}} = -\partial_\mu \phi^* \partial^\mu \phi - i \bar{\xi} \bar{\sigma}^\mu \partial_\mu \xi, \quad (10.111)$$

i.e. a system of massless bosons and fermions. Supersymmetry can be simply represented for this system by writing $Q_\alpha \phi \propto \xi_\alpha$, $\bar{Q}_{\dot{\alpha}} \xi_\beta \propto (\sigma^\mu)_{\dot{\alpha}\beta} \partial_\mu \phi$, as well as $Q_\alpha \xi_\beta = 0$ and $\bar{Q}_{\dot{\alpha}} \phi = 0$.

The Lagrangian (10.111) must respect the supersymmetry, i.e. we want it to be invariant under transformations that exchange the roles of ϕ and ξ . Let us assume, moreover, that we are interested in an infinitesimal transformation. Assuming the simplest possible transformation, the scalar field transforms according to $\delta_\zeta \phi = A \zeta^\alpha \xi_\alpha + \bar{A} \bar{\zeta}^{\dot{\alpha}} \bar{\xi}_{\dot{\alpha}}$, with A and \bar{A} constant. Since ϕ has dimension of mass, i.e. $[\phi] = M$ and $[\xi] = M^{3/2}$, one has $[\zeta] = M^{-1/2}$. As for the fermion, it must give a term involving the scalar field, and since $[\xi \phi] = M^{1/2}$, in the absence of a dimensionful constant in the theory (10.111), this term should necessarily be proportional to $\partial_\mu \phi$. Then, taking into account the fermionic index of ξ_α , we find the unique possible transformation: $\delta_\zeta \xi_\alpha = B \bar{\xi}^{\dot{\alpha}} (\sigma^\mu)_{\alpha\dot{\alpha}} \partial_\mu \phi$, with B a numerical constant.

We now require that these transformations represent effectively the supersymmetry algebra. For this, we note that to ensure (10.110), which can be written as $[\delta_\zeta, \delta_\eta] \phi = 2i(\eta \sigma^\mu \bar{\zeta} - \zeta \sigma^\mu \bar{\eta}) \partial_\mu \phi$, we should impose $\bar{A} = 0$ (no term in $\partial_\mu \phi^*$) and $AB = 2i$. Respecting the supersymmetry algebra on the spinor requires $[\delta_\zeta, \delta_\eta] \xi_\alpha = 2i(\eta \sigma^\mu \bar{\zeta} - \zeta \sigma^\mu \bar{\zeta}) \partial_\mu \xi_\alpha$, which is verified for the same constraints as long as the spinor also satisfies its equation of motion $\bar{\sigma}^\mu \partial_\mu \xi = 0$.

Varying the Lagrangian (10.111), we find a term that can be expressed as a total derivative, as expected if the invariance is respected, as long as we also have $A = iB^*$, so that we finally need to have $\delta_\zeta \phi = \sqrt{2} \zeta^\alpha \xi_\alpha$ and $\delta_\zeta \xi_\alpha = -i\sqrt{2} \bar{\xi}^{\dot{\alpha}} (\sigma^\mu)_{\alpha\dot{\alpha}} \partial_\mu \phi$.

These transformations do not actually indicate the existence of an arbitrary symmetry since, in order to obtain it, we impose that the spinor is on the mass shell (satisfying $p^2 + m^2 = 0$), which can be seen as an additional condition. We can notice that this condition does not apply for the scalar field.

10.3.2.2 Auxiliary field

The Lagrangian (10.111) cannot be invariant independently from the spinor equation of motion: as long as this equation is not satisfied, the bi-spinor has two complex functions, that is four degrees of freedom, to be compared with the two from the complex scalar field ϕ . To restore the parity between bosonic and fermionic degrees of freedom, we should thus add a complex scalar field, which is called an auxiliary field and is traditionally denoted by F . Having said that, this field is not a real physically observable field since supersymmetry can only be achieved provided ξ satisfies its equation of motion.

The supersymmetric model is thus the following:

$$\mathcal{L}_{wz} = -\partial_\mu \phi^* \partial^\mu \phi - i \bar{\xi} \bar{\sigma}^\mu \partial_\mu \xi + F^* F, \quad (10.112)$$

and the invariance is that with respect to the transformations

$$\delta_\zeta \phi = \sqrt{2} \zeta \xi, \quad \delta_\zeta \xi = \sqrt{2} \zeta F - i\sqrt{2} \sigma^\mu \bar{\zeta} \partial_\mu \phi \quad \text{and} \quad \delta_\zeta F = i\sqrt{2} \partial_\mu \xi \sigma^\mu \bar{\zeta}. \quad (10.113)$$

One can check that these transformations induce a variation $\delta_\zeta \mathcal{L}_{\text{wz}}$ that is a total derivative without using the equations of motion. This also translates into the fact that the additional F term closes the algebra.

From (10.112) we immediately see that the field F is essentially not physical. In fact, its dynamical equation simply implies $F = 0$ on its mass shell, and so it is only a constraint. Moreover, (10.112) and (10.113) show that $[F] = M^2$, unlike a normal scalar field.

10.3.2.3 Superfields

It is possible to construct an object that includes all the components of a supersymmetry multiplet, this is known as a superfield. In the case of a chiral theory (10.112), we simply write $\Phi = (\phi, \xi, F)$, so that the chiral superfield Φ includes both the physical fields ϕ and ξ as well as the auxiliary field F . The transformation laws of a chiral superfield are given by (10.113).

To add two superfields, it is sufficient to add their components, so that, since the transformations (10.113) are linear (as they are infinitesimal), the combined components transform as the sums of the transformed components.

The multiplication is more subtle. We proceed as follows: considering the two chiral superfields $\Phi_1 = (\phi_1, \xi_1, F_1)$ and $\Phi_2 = (\phi_2, \xi_2, F_2)$. We define the scalar part of the product $\phi = \Phi_1 \Phi_2|_{\text{scalar}} \equiv \phi_1 \phi_2$. Applying the transformation law to this scalar part, we see that it is necessary to have as spinor components $\xi = \Phi_1 \Phi_2|_{\text{spinor}} = \xi_1 \phi_2 + \xi_2 \phi_1$, whose transformation law allows us to obtain the composition law of the F terms, namely⁴ $F = \Phi_1 \Phi_2|_F = F_1 \phi_2 + F_2 \phi_1 - \xi_1 \xi_2$. In other words, we obtain

$$\boxed{\Phi_i \Phi_j = (\phi_i \phi_j, \xi_i \phi_j + \xi_j \phi_i, \phi_i F_j + \phi_j F_i - \xi_i \xi_j)}, \quad (10.114)$$

which generalizes the previous considerations to the case of an arbitrary number of superfields.

There is a formalism, based on the notion of *superspace*, composed of the four usual space-time coordinates x^μ , i.e. that commute, to which we add four anticommuting fermionic coordinates θ^α . In this formalism, the transformation laws such as (10.114) arise naturally from the Taylor series of any function of the coordinates [31].

10.3.2.4 Interactions, superpotential and F term

The transformation law of the F term in (10.113) suggests the use of such terms to automatically introduce Lagrangians that are invariant under a supersymmetry transformation: indeed, as soon as we consider *global supersymmetries*, i.e. for which the infinitesimal parameter ζ is a constant in space and time, then $\delta_\zeta F = \partial_\mu (i\sqrt{2}\xi^\mu \bar{\zeta})$ and so it is a total derivative. We are thus certain of having the required invariance as soon as we take a theory with Lagrangian

$$\mathcal{L}_{\text{chiral}} = - \sum_i (\partial_\mu \phi_i^\star \partial^\mu \phi_i + i \bar{\xi}_i \bar{\sigma}^\mu \partial_\mu \xi_i + F_i^\star F_i) + W(\Phi_i)|_F + [W(\Phi_i)|_F]^*, \quad (10.115)$$

⁴To obtain this relation, we should go back to the definition of the scalar product of two spinors and use the relation $\epsilon^{\beta\rho} \epsilon_{\alpha\sigma} = \frac{1}{2} \delta_\alpha^\beta \delta_\sigma^\rho$, while remembering that both spinor components anticommute.

where W , which is known as the *superpotential*, is a function of the superfields,⁵ and is a priori arbitrary but is in practice constrained so that the theory is renormalizable. Note that in (10.115), we only include in the Lagrangian the F term corresponding to the superpotential (and its complex conjugate such that the resulting Lagrangian is real): this is the meaning of the notation $|_F$.

This superpotential allows for the construction of a scalar potential and all the interaction terms between the different fields contained in the theory. For a single multiplet, the most commonly considered case is

$$W(\Phi) = \frac{1}{2}m\Phi\Phi + \frac{1}{6}g\Phi\Phi\Phi, \quad (10.116)$$

for which the F term is given by

$$W(\Phi)|_F = m\phi F - \frac{1}{2}m\xi\xi + \frac{g}{2}(\phi^2 F - \phi\xi\xi),$$

and after some computation we find

$$\begin{aligned} \mathcal{L} = & -\partial_\mu\phi^*\partial^\mu\phi - m^2\phi^2 - i\bar{\xi}\bar{\sigma}^\mu\partial_\mu\xi - \frac{1}{2}m(\xi\xi + \bar{\xi}\bar{\xi}) + \tilde{F}^*\tilde{F} \\ & - \frac{1}{2}\phi\xi\xi - \frac{1}{2}\phi^*\bar{\xi}\bar{\xi} - gm|\phi|^2\Re(\phi) - \frac{g^2}{4}|\phi|^4, \end{aligned} \quad (10.117)$$

with $\tilde{F} = F + m\phi^* + \frac{1}{2}g\phi^{*2}$ a new auxiliary field that is required to be non-dynamical since its equation of motion is still $\tilde{F} = 0$ (it is actually obtained by setting $\tilde{F} = \delta\mathcal{L}/\delta F^*$).

This model of a chiral supermultiplet thus contains a left fermion and an interacting scalar field, both of them of the same mass m . Actually, it even turns out that due to the precise form of the interaction terms, which is entirely due to the fact that the theory is supersymmetric, the masses of these two particles are exactly equal to all orders in perturbations. This is an example of a *non-renormalization* property, which generally arise in supersymmetric theories.

10.3.2.5 Scalar potential

Starting from the superpotential W , and from the construction of its F term, the scalar field potential that we find, for instance, in (10.117) is very simply obtained thanks to the following relation:

$$V(\phi_i) = \sum_i \left| \frac{\partial W}{\partial \phi_i} \right|^2, \quad (10.118)$$

in which relation we have used the general form with an arbitrary number of fields, but that immediately gives all the terms not involving the scalar parts of the action (10.117), when we restrict ourselves to the case of a single multiplet.

⁵Notice that as a function of the scalar components, the superpotential depends of the fields ϕ_i but cannot depend on their conjugate ϕ^{*j} .

The scalar potential of (10.118) comes directly from the F terms of the initial Lagrangian. Note that the quantities F_i are solutions of the constraint equations for these auxiliary quantities

$$F_i^* = -\frac{\partial W}{\partial \phi_i}, \quad (10.119)$$

so that $V = \sum_i |F_i|^2$.

10.3.3 Gauge field

Interactions between particles, besides the ones specific to a scalar field, are propagated by gauge fields. If the theory that correctly describes Nature is supersymmetric, then these gauge fields must also belong to supermultiplets, which we thus call gauge or vector multiplets. We have seen that the Haag–Łopuszański–Sohnius theorem requires the generator of the supersymmetry to be fermionic, and of spin $\frac{1}{2}$. To construct a supermultiplet containing a vector field, it is thus easiest to consider fermionic superpartners of spin $\frac{1}{2}$.

10.3.3.1 Gauge invariance

As soon as a theory contains gauge fields $C_{a\mu}$ that are in the adjoint representation of an invariance group, in order to be made supersymmetric it must contain a set of bi-spinors λ_a [not to be confused with the elements of the representation satisfying (2.119) of Chapter 2, that we will denote here by T^a]; we will therefore have $[T^a, T^b] = i\Gamma^{ab}{}_c T^c$, and the representation satisfies (2.121), i.e. the elements of the T^a matrices are $(T^b)_c{}^a = i\Gamma^{ab}{}_c$.

Due to this representation, the covariant derivative of these fermions is

$$(D_\mu \lambda)^a = (\partial_\mu - igT^b C_{b\mu}) \chi^a = \partial_\mu \chi^a + g\Gamma^{ab}{}_c C_{b\mu} \lambda^c,$$

and the gauge invariance thus requires that the Lagrangian is invariant under the transformations

$$\begin{cases} C_{a\mu} \mapsto C'_{a\mu} = C_{a\mu} - \frac{1}{g}\partial_\mu \alpha_a + \Gamma^{bc}{}_a \alpha_b C_{c\mu}, \\ \lambda_a \mapsto \lambda'_a = \lambda_a + \Gamma^{bc}{}_a \alpha_b \lambda_c. \end{cases}$$

Proceeding analogously to the chiral case, we find that the supersymmetric transformation for the gauge field must be

$$\delta_\zeta C_a^\mu = i(\bar{\zeta} \bar{\sigma}^\mu \lambda_a - \bar{\lambda}_a \bar{\sigma}^\mu \zeta), \quad (10.120)$$

where we have chosen the representation so that $Q_\alpha C_{a\mu} = 0$.

10.3.3.2 D term

Just as we needed to associate a bi-spinor to a complex scalar field having two components, for a gauge field C_μ that has 4 components ($\mu = 0, \dots, 3$), two bi-spinors are, in general, required. As for the chiral multiplet, the bi-spinor components are complex, so that there are as many additional spinor degrees of freedom: now, we need several auxiliary fields that we denote by D_a .

We then find that the transformation laws required to make the theory supersymmetric are

$$\delta_\zeta \lambda_a = \frac{1}{2} \sigma^\mu \bar{\sigma}^\nu \zeta F_{a\mu\nu} + i\zeta D_a \quad \text{and} \quad \delta_\zeta D_a = \bar{\zeta} \bar{\sigma}^\mu D_\mu \lambda_a + D_\mu \bar{\lambda}_a \bar{\sigma}^\mu \zeta, \quad (10.121)$$

with $F_{a\mu\nu} \equiv \partial_\mu C_{a\nu} - \partial_\nu C_{a\mu} + g \Gamma^{bc}{}_a C_{b\mu} C_{c\nu}$.

The model that implements this symmetry can now be written as:

$$\mathcal{L}_{\text{vector}} = \sum_a \left(-\frac{1}{4} F_{a\mu\nu} F_a^{\mu\nu} - \bar{\lambda}_a \bar{\sigma}^\mu D_\mu \lambda_a + \frac{1}{2} D_a D_a \right). \quad (10.122)$$

10.3.3.3 D term of the potential

A complete theory contains chiral as well as vector multiplets. This means that the kinetic terms in (10.115) must take the gauge invariance into account, which is achieved by replacing the partial derivative ∂_μ by covariant derivatives D_μ , for both the scalar fields and their partners. This induces two consequences in the transformation law (10.113) of the F term: the first one is that the partial derivative must be replaced by a covariant derivative, but furthermore, the interaction between ϕ and $C_{a\mu}$ introduces an additional term. We find

$$\delta_\zeta F = i\sqrt{2} D_\mu \xi \sigma^\mu \zeta + 2ig T^a \phi \bar{\zeta} \bar{\lambda}_a, \quad (10.123)$$

as well as new interaction terms, namely

$$\mathcal{L}_{\text{int}} = \sqrt{2} gi [(\phi^* T^a \xi) \lambda_a - \bar{\lambda}_a (\bar{\xi} T^a \phi)] + g (\phi^* T^a \phi) D_a. \quad (10.124)$$

As seen in (10.124), we find a new component for the self-coupling potential of the scalar field, which is finally written as

$$V(\phi_i) = \sum_i |F_i|^2 + \frac{1}{2} \sum_a |D_a|^2 = \sum_i \left| \frac{\partial W}{\partial \phi_i} \right|^2 + \frac{g^2}{2} \sum_a (\phi^* T^a \phi)^2, \quad (10.125)$$

where, in the last equality, the D term has been replaced by its value as a function of the scalar field using its equation of motion, i.e. $D^a = -g \phi^* T^a \phi$. The effective potential is thus the sum of two terms, arising for the two chiral and vectorial sectors of the supersymmetric theory. We call them, respectively, F and D terms of the potential.

10.3.4 Supersymmetry breaking

In the framework of supersymmetry, the particles of the standard model described in Chapter 1 are supposed to belong to chiral supermultiplets for the quarks and leptons, and vectorial supermultiplets for the gauge fields. This means that if supersymmetry is realized in Nature, we must necessarily have scalar fields of the same mass as the quarks or leptons, called *squarks* and *sleptons*. Similarly, we should see fermionic superpartners of the gauge fields, the *gauginos*. All these particles are obviously not

observed, which does not mean that supersymmetry is not realized, but simply that if it is, it must be broken.

It is interesting to note that the scalar potential arising from the supersymmetric theory is a positive-definite function. As shown in (10.107), the vacuum is supersymmetric as long as the Hamiltonian has zero value in this state to respect the supersymmetry the potential itself should vanish, and so both terms, F and D , should be positive, separately. This leads to two ways to break the supersymmetry.

10.3.4.1 Soft SUSY breaking terms

The simplest way to break supersymmetry amounts to explicitly introducing terms that do not satisfy it. These terms, as long as they appear at low energy scales, make it possible to maintain the useful properties of the theory, in particular the resolution of the hierarchy problem without resorting to a fine tuning of the parameters at each order in perturbation, and are called the soft SUSY breaking terms. They are the following:

$$\mathcal{L}_{\text{soft}} = -\frac{1}{2} \sum_a m_\lambda^a \lambda_a \lambda_a - \frac{1}{2} m_{(ij)}^2 \phi_i \phi_j^* - \frac{1}{2} B^{ij} \phi_i \phi_j - \frac{1}{6} A^{ijk} \phi_i \phi_j \phi_k + \text{h.c.}, \quad (10.126)$$

i.e. specific masses for the gauginos and the scalars, as well as bi- and trilinear couplings for the scalar fields. Although seemingly arbitrary, these terms can naturally arise in more extended theories such as supergravity and superstrings [34].

10.3.4.2 Breaking by an F term

The first way to effectively implement supersymmetry breaking amounts to performing a spontaneous breaking, of the Higgs-model type. For a chiral multiplet, the only solution is to have $\langle 0|F|0 \rangle \neq 0$.

The potential can have several minima, and in order for the supersymmetry to be effectively broken, it is necessary that none of them is supersymmetric. For the vacuum to be supersymmetric, we need $V = 0$, and this defines the absolute minimum of the energy, which is positive definite. This considerably restricts the possible superpotentials allowing for supersymmetry breaking: for instance, if there are nonlinear terms in the fields, then (10.119) are homogeneous, and always admit the solution $\{F_i = 0, \forall i\}$. We are then ensured that this configuration is indeed the absolute minimum of the theory.

The superpotential (10.116) can be generalized to several fields with linear terms. This gives

$$W(\Phi_i) = \sum_i \alpha_i \Phi_i + \frac{1}{2} \sum_{ij} m_{ij} \Phi_i \Phi_j + \frac{1}{6} \sum_{ijk} \lambda_{ijk} \Phi_i \Phi_j \Phi_k, \quad (10.127)$$

leading to a renormalizable potential.

A model based on the superpotential (10.127) has been proposed by O’Raifeartaigh [35], having three chiral superfields Φ_1 , Φ_2 and Φ_3 whose dynamics are governed by

$$W(\Phi_1, \Phi_2, \Phi_3) = \alpha \Phi_1 (\Phi_2^2 - \eta^2) + m \Phi_2 \Phi_3, \quad (10.128)$$

whose F terms are, respectively, $F_1^* = -\alpha(\phi_2^2 - \eta^2)$, $F_2^* = -2\alpha\phi_1\phi_2 + m\phi_3$ and $F_3^* = -m\phi_2^2$. It is clear that simultaneously cancelling F_1 and F_3 is impossible since the first one requires $\phi_2 = \eta$ and the second $\phi_2 = 0$. Consequently, the global F term does not vanish, so that the supersymmetry is spontaneously broken.

The potential we obtain from (10.128) is

$$V = \alpha^2 |\phi_2^2 - \eta^2|^2 + |2\alpha\phi_1\phi_2 + m\phi_3|^2 + m^2 |\phi_2|^2, \quad (10.129)$$

whose absolute minimum, if $\alpha\eta > m$, is at $\phi_2 = \phi_3 = 0$. This minimum is reached independent of the value of ϕ_1 : the potential is said to have a flat direction for ϕ_1 . We then find that the potential takes the constant value $V = \alpha^2\eta^4 \neq 0$.

The supersymmetry breaking appears even more clearly when studying the particle mass spectrum of these models, which follows from the complete expression for the Lagrangian in which we write that the scalar fields take their expectation value. We find that ξ_1 , the fermion associated to ϕ_1 , itself in the flat direction, is massless: it is a Goldstone boson, similar to the boson of the same name in the models of symmetry breaking. Choosing $\langle\phi_1\rangle = 0$ for the sake of simplicity (but this is a physical choice, since the fermion masses effectively depend on this choice), then we find that the spinors ξ_2 and ξ_3 are both of mass m .

As for the masses of the scalar particles, it is sufficient to replace the fields ϕ_i by their expectation values (here, zero for all fields) and to look at the resulting potential for perturbations around these expectation values. We find that ϕ_1 is massless, and that ϕ_3 , like its partner ξ_3 , is of mass m : the theory remains supersymmetric from the point of view of these fields. On the other hand, expanding ϕ_2 in real and imaginary parts as

$$\phi_2 = \frac{1}{\sqrt{2}} [\Re(\phi_2) + i\Im(\phi_2)] \equiv \frac{1}{\sqrt{2}} (\phi_{\Re} + i\phi_{\Im}),$$

we find a quadratic term

$$V_{(2)} = \frac{1}{2} (m^2 - 2\alpha^2\eta^2) \phi_{\Re}^2 + \frac{1}{2} (m^2 + 2\alpha^2\eta^2) \phi_{\Im}^2 + \frac{1}{2} (m_{\Re}^2 \phi_{\Re}^2 + m_{\Im}^2 \phi_{\Im}^2),$$

so that the masses of the real and imaginary parts of the scalar ϕ_2 are different, and different from that of the associated fermion ξ_2 : supersymmetry is effectively broken. We reach the same conclusion independent of the value chosen for $\langle\phi_1\rangle$. Moreover, we can notice a more general phenomena, namely that there is an apparent breaking for the multiplet Φ_2 that, in the superpotential, turns out to couple to Φ_1 , i.e. the one containing the Goldstone boson.

10.3.4.3 Breaking by a D term

There is another way [36] to break supersymmetry, which amounts to observing that the variation of the D term, given by (10.121), reduces to a total derivative as long as the invariance group is $U(1)$: in this case, the additional terms in the covariant derivatives change sign for λ and $\bar{\lambda}$ since one is the charge conjugate of the other, and we then obtain $\delta_\zeta D = \partial_\mu (\bar{\zeta}\bar{\sigma}^\mu\lambda + \bar{\lambda}\bar{\sigma}^\mu\zeta)$. Notice that for a non-Abelian symmetry, the generator signs are interchanged since λ^a and $\bar{\lambda}^a$ are in the same group representation,

and the supersymmetric variation of D^a gives a non-derivative term, proportional to the structure constants of the group.

In the case of a U(1) invariance, it is now clear that we can add a term proportional to D in the Lagrangian, called a Fayet–Iliopoulos term

$$\mathcal{L}_{\text{FI}} = \kappa D, \quad (10.130)$$

so that the ‘ D ’ part of the total Lagrangian is

$$\mathcal{L}_D = \frac{1}{2} D^2 + \kappa D + g \sum_i (q_i |\phi_i|^2) D, \quad (10.131)$$

where the q_i are the charges of the scalar fields ϕ_i , under the U(1) in question. We thus obtain, upon varying with respect to D ,

$$D = - \left[\kappa + g \sum_i (q_i |\phi_i|^2) \right],$$

giving a D term to the potential

$$V_D = -\mathcal{L}_D = \frac{1}{2} D^2 = \frac{1}{2} \left[\kappa + g \sum_i (q_i |\phi_i|^2) \right]^2, \quad (10.132)$$

so that if the symmetry is not broken, i.e. if $\phi_i = 0, \forall i$, we then find $D = -\kappa$, and the total potential (10.125) is $V = \frac{1}{2}\kappa^2 > 0$, indicating that the supersymmetry is indeed broken.

The advantage of this method is that it offers the possibility to break SUSY with only one chiral multiplet, unlike the F term. The supersymmetry breaking is easily verified by studying this specific case, with a single chiral multiplet of charge q : if $gq\kappa > 0$, we must have $\phi = 0$ to minimize the potential, and we find a mass term for ϕ with $m_\phi^2 = gq\kappa$ whereas, at the same time, we notice that the fermion remains massless.

The inconvenience is that it requires a symmetry under the transformations of a U(1) group and, for phenomenological reasons, it turns out that we cannot use the one of the standard model U(1) _{γ} .

10.3.5 The minimal supersymmetric standard model (MSSM)

Since no pairs of particles from the standard model can be used to form a supermultiplet, if we want to create a supersymmetric version of this model, it is necessary to duplicate all the known particles. This leads to a model, already introduced by Fayet in 1976 [37] called the minimal supersymmetric standard model (MSSM), containing a large number of particles. This model is described in detail in Refs. [38, 39], and we sketch here its main features.

10.3.5.1 Chiral multiplets

Each quark and lepton appears in the form of a left or right fermion, the formalism described so far is particularly well adapted since it is sufficient to write the right

spinors in the form of left spinors [see (10.102)]. We then place all these fermions in chiral multiplets. The lepton partners are called *sleptons*, the quark ones *squarks*, and the Higgs ones *Higgsinos*: the general nomenclature is to construct the name of the bosons associated to the fermions of the standard model by adding an ‘s’ (for ‘scalar’) before the particle name, whereas to the partner of a boson is attributed the name to which we add the suffix ‘ino’. These names, as well as the usual notations, are summarized in Table 10.1.

Table 10.1 Chiral supermultiplets in the minimal supersymmetric standard model.

| Particle names | | spin 0 | spin $\frac{1}{2}$ |
|---|-----------|-----------------------------|-----------------------------------|
| squarks, quarks ($\times 3$ families) | Q | $(\bar{u}_L \ \tilde{d}_L)$ | $(u_L \ d_L)$ |
| | \bar{u} | \tilde{u}_R^* | u_R^\dagger |
| | \bar{d} | \tilde{d}_R^* | d_R^\dagger |
| sleptons, leptons ($\times 3$ families) | L | $(\bar{\nu} \ \tilde{e}_L)$ | $(\nu \ e_L)$ |
| | \bar{e} | \tilde{e}_R^* | e_R^\dagger |
| Higgs, higgsinos | H_u | $(H_u^+ \ H_u^0)$ | $(\tilde{H}_u^+ \ \tilde{H}_u^0)$ |
| | H_d | $(H_d^0 \ H_d^-)$ | $(\tilde{H}_d^0 \ \tilde{H}_d^-)$ |

The phenomenology of the particles we know, and in particular the fact that a mass should be given to the quarks of type ‘u’, imposes that no slepton can play the role of the Higgs boson. It is thus necessary to include a multiplet H_u containing the Higgs to the model. In this same multiplet, we find an electrically charged scalar field. The associated fermion \tilde{H}_u must therefore be a Dirac fermion since a Majorana fermion does not conserve the electric charge. As a result, a second bi-spinor \tilde{H}_d should be introduced, which is different from the first one and forms with it a Dirac quadrispinor. This second Higgsino then requires the existence of a second Higgs: the MSSM has an additional Higgs field.

The superpotential we consider is then

$$W_{\text{SM}} = \epsilon_{ij} \left(y_e H_d^i L^j e^c + y_d H_d^i Q^j d^c + y_u H_u^i Q^j u^c + \mu H_u^i H_d^j \right), \quad (10.133)$$

for each generation, i.e. there must be three superpotentials of this kind to reproduce all the particles we know. The Yukawa couplings y_e , y_u and y_d can also be seen as matrices in the space of generations. The constant μ is required to ensure that all the Higgs bosons are massive.

In this framework, the Higgs mechanism of Chapter 2 is implemented thanks to the two Higgs fields. We find

$$\langle H_d \rangle = \begin{pmatrix} v_d \\ 0 \end{pmatrix} \quad \text{and} \quad \langle H_u \rangle = \begin{pmatrix} 0 \\ v_u \end{pmatrix}, \quad (10.134)$$

and we recover the vacuum expectation value η of (2.162) by writing $\eta^2 = v_d^2 + v_u^2$. We then define a parameter β giving the ratio of the expectation values

$$\tan \beta \equiv \frac{v_u}{v_d}.$$

(10.135)

The constraints obtained on supersymmetric dark-matter candidates are often expressed in terms of this parameter.

This part of the MSSM is completed by soft breaking terms, in particular the masses m_u and m_d for the Higgs multiplets.

10.3.5.2 Charginos and neutralinos

The two charged fermions associated to the Higgs field, the two higgsinos, belong to the more general category of ‘charginos’. The other particles of this class are associated to the charged gauge bosons W^\pm , they are denoted by \widetilde{W}^\pm and are called ‘charged gauginos’.

The two other gauge bosons of the standard model of Chapter 2 are neutral, and their superpartners are called ‘neutralinos’. This category is completed by the two neutral higgsinos, which are also fermions. It is common in the context of the MSSM model to denote by W_μ^i the gauge fields of $SU(2)_L$ [B_μ^i in (2.154)], and by B_μ that of $U(1)_Y$ [C_μ in (2.154)]. Due to this, the partners are called, respectively, the ‘winos’ and ‘binos’. Finally, on the basis of the states defined by (2.158) and (2.159), we can also define the photino $\widetilde{\gamma}$ and the zino \widetilde{Z} , the partners of, respectively, the photon and the Z^0 , with, for instance,

$$\widetilde{\gamma} = \widetilde{W}^3 \sin \theta_w + \widetilde{B} \cos \theta_w,$$

but this nomenclature is not used very much: in regions with energies of interest from the viewpoint of supersymmetry, the latter can be broken or not, whereas the gauge invariance, on the other hand, is not broken so that it is actually the gauge eigenstates, i.e. W^i and B , which are useful

Table 10.2 Vector supermultiplets in the minimal supersymmetric standard model.

| Particle names | spin $\frac{1}{2}$ | spin 1 |
|-----------------|---------------------------------------|---------------|
| gluino, gluon | \widetilde{g} | g |
| winos, W bosons | \widetilde{W}^\pm \widetilde{W}^0 | W^\pm W^0 |
| bino, B boson | \widetilde{B}^0 | B^0 |
| photino, photon | $\widetilde{\gamma}$ | γ |

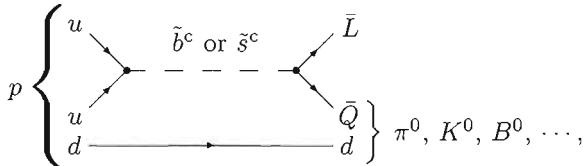
10.3.5.3 Proton decay and R-parity

The superpotential (10.133) is not the only one we can accommodate in the MSSM that satisfies all the symmetries. In particular, we can imagine terms that explicitly violate the conservation of lepton and baryon numbers, which arise by accident in the standard model. Such a term is given by

$$W_{LB} = \lambda^{ijk} L_i L_j e_k^c + \lambda' ijk L_i Q_j d_k^c + \lambda'' ijk u_i^c d_j^c d_k^c + \mu'^i L_i H_u, \quad (10.136)$$

and leads to new interactions that are very constrained by experiments.

By combining these interactions, we find that the following decay channel for the proton is possible,



where the virtual particles exchanged do not necessarily have extraordinary high masses (of the order of the TeV so that the SUSY breaking allows the unification of Fig. 10.5). Taking into account these kinds of diagram, the proton lifetime is considerably shorter than the current limits, by many orders of magnitudes, and these interactions should thus be suppressed.

A good justification [37] to forbid the terms (10.136) of the MSSM superpotential is to postulate the existence of a discrete symmetry indicating if a particle is a particle initially in the standard model or if it is a partner. Writing

$$R = (-1)^{3B+L+2s}, \quad (10.137)$$

where B is the baryonic number ($\frac{1}{3}$ for each quark, zero for the leptons and photons), L the leptonic number (-1 for the electron, zero for the quarks and photon) and s the spin ($\frac{1}{2}$ for the fermions, zero for the scalars, and 1 for the gauge fields), then we find that all the particles of the standard model have $R = +1$, which implies that since the spin is modified by $\frac{1}{2}$, that the partners all have $R = -1$. As a consequence, this is a Z_2 symmetry that we should impose.

10.3.5.4 The lightest supersymmetric partner (LSP)

The R -parity has several consequences, in particular in cosmology, where we have seen, in Chapter 11, that such an unbroken discrete invariance between the particles implies the existence of cosmic strings in the Universe.

Another consequence is the existence of a stable particle of large mass, which is the lightest supersymmetric particle: since we impose that no transition can violate R -parity, all the supersymmetric particles can in the end decay into the lightest one, but this one is stable. This particle, called ‘LSP’, is a very serious candidate for dark matter, see §. 7.2.5.2 of Chapter 7. In its minimal version, the most general supersymmetry model has 124 free parameters, so that we do not know the mass values of all the superparticles, and as a consequence, the identity of the LSP is not a priori determined.

10.3.6 Supergravity

So far, we have only considered theories invariant under *global* supersymmetry transformations, i.e. the ones for which the parameter ζ of the transformation does not depend on the position. Moreover, we also explicitly required these theories to be renormalizable, which considerably constrains the form of the superpotential, but also suppresses some terms that are otherwise possible, such as the Kähler potential discussed below.

10.3.6.1 The gravitino

When requiring the supersymmetry to be gauged, i.e. that the transformation is now made through a space-time-dependent spinor $\zeta(x)$, many of the previous results are modified. We have seen that from (10.105), two SUSY transformations were essentially equivalent to a translation. If one demands invariance under local supersymmetric transformations, one then necessarily includes different translations at each point of space-time: they are general transformations, isomorphisms of space-time itself. As a consequence, the invariance we are describing is that of general relativity, which is thus naturally incorporated into the framework of particle physics. This is why local supersymmetry is called *supergravity*, often denoted by SUGRA [40].

Another way to see this point is that, in the limit where the Planck mass tends towards infinity, supergravity becomes global SUSY. Up to now, we have studied only renormalizable models, but since supergravity is already not renormalizable, it is possible to consider supersymmetric models that are not renormalizable either, where we understand that all these theories are seen as low-energy approximations, in particular to superstring theory.

As soon as we allow for gauged supersymmetric transformations, the terms proportional to $\partial_\mu \zeta(x)$, which will necessarily appear in transformations of the kinetic terms, must then be eliminated. This can only be done by introducing a field, equivalent to the gauge bosons, with not only a spinor index to compensate that of ζ , but also a vector index for ∂_μ : we expect an object Ψ_μ , whose transformation law is $\delta\zeta \psi \propto \partial_\mu \zeta$. In this case, it is a particle of spin $\frac{3}{2}$, which we call the *gravitino*, whose kinetic term is described by the Rarita–Schwinger action (see Section 7.2.5.2 of Chapter 7)

$$\mathcal{S}_{\text{RS}} = -\frac{1}{2} \int d^4x \epsilon^{\mu\nu\alpha\beta} \bar{\Psi}_\mu \gamma_5 \gamma_\nu \partial_\alpha \Psi_\beta. \quad (10.138)$$

It turns out that this gravitino must be placed in a supermultiplet containing a particle of spin 2, the graviton.

Despite what we might naively think, supergravity corrections are not always negligible, even when considering low-energy physics. For instance, in theories where supersymmetry is broken with an F term, the scale at which this happens can be of the order of $M_S \sim 10^{11}$ GeV, so that some particles obtain masses much less than this scale, particles of the standard model and their superpartners, but also other particles, arising from the supergravity sector, obtain masses of the order of M_S , so that the SUGRA corrections are to be taken into account. Similarly, in other situations and in particular in cosmology, some scalar fields take expectation values of the order of the Planck mass; here again, one cannot neglect corrections from supergravity.

10.3.6.2 Kähler potential

Unlike the case of global supersymmetry, it is not necessary to require that the SUGRA Lagrangian be renormalizable since the theory contains gravity and thus is not renormalizable (in four dimensions). Note, by the way, that as a consequence, we can also discard this constraint for global SUSY theories. We then obtain all the possible terms for a global supersymmetry compatible with supergravity. These terms represent what we call supergravity corrections.

The most general Lagrangian in this framework for a chiral multiplet Φ then contains

$$\mathcal{L}_{\text{kin}} = K^i_j \partial_\mu \phi_i \partial^\mu \phi^{*j}, \quad \text{where} \quad K^i \equiv \frac{\partial K}{\partial \phi_i} \quad K_j \equiv \frac{\partial K}{\partial \phi^{*j}} \quad \text{and} \quad K^i_j \equiv \frac{\partial^2 K}{\partial \phi_i \partial \phi^{*j}}, \quad (10.139)$$

and the *Kähler potential* K is an arbitrary function of the scalar fields ϕ_i and their conjugate ϕ^{*j} . This function, which is equivalent to a metric on the space of the scalar fields, must be dimensionless. Notice furthermore that we can add a gauge coupling matrix, $f_{ab}(\Phi_i)$, whose elements depend on the chiral superfields Φ_i that serves as a metric, in the same way as the Kähler coupling, for the kinetic terms of the gauge fields (the indices a, b are indices of the gauge group). In general, and in particular in what follows, we set $f_{ab} = \delta_{ab}$. Actually, if this function is non-trivial, then the coupling constants must vary with time in cosmology [3].

Once we have the Kähler potential, we find the scalar potential through the relation [31, 39]

$$V = \frac{e^K}{\kappa^2} \left[K^i (K^{-1})_i^j K_j - 3 \right], \quad (10.140)$$

where $(K^{-1})_i^j$ is the inverse of K^i_j , i.e. $(K^{-1})_i^j K^i_n = \delta_n^j$. The coefficient $\kappa = 8\pi/M_p^2 = 8\pi G_N$ is justified to make K dimensionless, knowing that in supergravity, the only energy scale at our disposal is the Planck mass.

10.3.6.3 $N = 1$ minimal supergravity

To conclude these extensions of the standard model, here is the special case of the so-called $N = 1$ minimal supergravity that illustrates the computation of the scalar potential in a precise case. In that case, we have

$$K = \kappa \phi_i \phi^{*i} + \ln (\kappa^3 |W|^2), \quad (10.141)$$

where W is the superpotential. With this Kähler potential, we find

$$K_j = \kappa \phi_j + \frac{1}{W^*} \frac{\partial W^*}{\partial \phi^{*j}}, \quad K^i = \kappa^2 \phi^{*i} + \frac{1}{W} \frac{\partial W}{\partial \phi_i},$$

and

$$K^i_j = \kappa \delta_j^i \implies (K^{-1})_j^i = \kappa^{-1} \delta_j^i,$$

so that applying (10.140) leads to

$$V(\phi, \phi^*) = e^{\kappa \phi_i \phi^{*i}} \left(\left| \frac{\partial W}{\partial \phi_i} + \kappa \phi^{*i} W \right|^2 - 3\kappa |W|^2 \right), \quad (10.142)$$

and we recover the usual relation $V = |\partial W / \partial \phi|^2$ in the limit $M_p \rightarrow \infty$, i.e. $\kappa \rightarrow 0$.

10.3.6.4 No-scale potential

Models of supergravity have, in general, a hidden sector, defined as the set of fields of the theory that only couple to ordinary matter (quarks, leptons, etc.) through gravitational interaction. In particular, the gravitino belongs to this sector. As seen earlier, the characteristic mass of the supersymmetry breaking should not be too different from 1 TeV to agree with grand unification theories. In supergravity models, this is not very natural since the only parameter with dimension of mass present in these theories is the Planck mass.

An interesting possibility is to resort to a flat potential, not depending on the fields (which in addition is interesting in terms of cosmology since it could allow for a natural slow-roll solution). This is, for instance, the case for the so-called *no-scale supergravity*' Kähler potentials, whose simplest example is

$$K_{\text{ns}} = -3 \ln(\phi + \phi^*), \quad (10.143)$$

giving an identically vanishing potential (10.140).

In this way, we can reasonably think that the masses of the particles associated to these fields will be produced through radiative corrections induced by the coupling with particles of non-vanishing expectation values, of the order of the Planck mass. These corrections are, in general, logarithmic in these VEV, so that the resulting mass can be naturally reduced exponentially. Other examples are discussed in Ref. [31].

The theoretical extensions discussed in this chapter are constantly used in cosmology, for several reasons. First, the construction of our cosmological model must rely on the most realistic physical theories and must thus include all their developments. For instance, since supersymmetry is thought to be necessary to build a consistent theory, one expects that supersymmetric grand unified models should be used to describe the first moments of the Universe. Furthermore, and this is probably the aspect we are the most interested in here, all these extensions lead to cosmological consequences that can be tested observationally. For instance, these can be the existence of topological defects such as those described in Chapter 11, or the possibility of implementing a phase of inflation without needing to add extra *ad-hoc* scalar fields (see Chapter 12). It follows that cosmological observations can help us constrain these extensions and possibly exclude some of them. Finally, these theoretical extensions are at the heart of string theory, which is briefly discussed in Chapter 13 and whose only eventual observable consequences are, at the present state of knowledge, in the area of cosmology.

References

- [1] P. JORDAN, ‘Formation of the stars and development of the Universe’, *Nature* **164**, 637, 1949; M. FIERZ, *Helv. Phys.* **29**, 128, 1956; P. JORDAN, ‘Zum gegenwärtigen Stand der Diracschen kosmologischen Hypothesen’, *Z. Phys.* **157**, 112, 1959; C. BRANS and R. DICKE, ‘Mach’s principle and a relativistic theory of gravitation’, *Phys. Rev.* **124**, 925, 1961.
- [2] T. DAMOUR and G. ESPOSITO-FARÈSE, ‘Tensor-multi scalar theories of gravitation’, *Class. Quant. Grav.* **9**, 2093, 1992.
- [3] J.-P. UZAN, ‘The fundamental constant and their variation: observational and theoretical status’, *Rev. Mod. Phys.* **75**, 403, 2002.
- [4] C. WILL, *Theory and experiments in gravitational physics*, Cambridge University Press, 1993.
- [5] G. ESPOSITO-FARÈSE and D. POLARSKI, ‘Scalar-tensor gravity in an accelerating Universe’, *Phys. Rev. D* **63**, 063504, 2001.
- [6] G. ESPOSITO-FARÈSE, ‘Binary-pulsar tests of strong-field gravity and gravitational radiation damping’, Proceedings of the 10th Marcel Grossmann Meeting, Word Scientific (2005) 647 [[arXiv:gr-qc/0402007](https://arxiv.org/abs/gr-qc/0402007)].
- [7] B. M. BARKER, ‘General scalar-tensor theory of gravity with constant G’, *Astrophys. J.* **219**, 5, 1978.
- [8] C. SCHIMD, J.-P. UZAN and A. RIAZUELO, ‘Weak lensing in scalar-tensor theories of gravity’, *Phys. Rev. D* **71**, 083512, 2005.
- [9] A. RIAZUELO and J.-P. UZAN, ‘Cosmological observations in scalar-tensor quintessence’, *Phys. Rev. D* **66**, 023525, 2002.
- [10] T. DAMOUR and K. NORDTVEDT, ‘General relativity as a cosmological attractor of tensor-scalar theories’, *Phys. Rev. Lett.* **70**, 2217, 1993.
- [11] T. DAMOUR and B. PICHON, ‘Big-Bang nucleosynthesis and tensor-scalar gravity’, *Phys. Rev. D* **59**, 123502, 1999.
- [12] A. COC, et al., ‘Big-Bang nucleosynthesis constraints on tensor-scalar theories of gravity’, *Phys. Rev. D* **73**, 083525, 2006.
- [13] J.-P. UZAN, ‘Cosmological scaling solutions of non-minimally coupled scalar fields’, *Phys. Rev. D* **59**, 123510, 1999.
- [14] M. GASPERINI, F. PIAZZA and G. VENEZIANO, ‘Quintessence as a run-away dilaton’, *Phys. Rev. D* **65**, 023508, 2002.
- [15] T. DAMOUR, F. PIAZZA and G. VENEZIANO, ‘Violations of the equivalence principle in a dilaton-runaway scenario’, *Phys. Rev. D* **66**, 081601, 2002.
- [16] N. BARTOLO and M. PIETRONI, ‘Scalar-Tensor Gravity and Quintessence’, *Phys. Rev. D* **61**, 023518, 2000.
- [17] D. WANDS, ‘Extended gravity theories and the Einstein–Hilbert Action’, *Class. Quant. Grav.* **5**, 269, 1994.

- [18] G. MANGANO and M. SOKOLOWSKI, ‘Physical equivalence between nonlinear gravity theories and a general self-gravitating scalar field’, *Phys. Rev. D* **50**, 5039, 1994.
- [19] T. JACOBSON, ‘Introduction to quantum fields in curved spacetime and the Hawking effect’, [gr-qc/0308048].
- [20] N. D. BIRRELL and P. C. DAVIES, *Quantum fields in curved space*, Cambridge University Press, 1984.
- [21] P. R. ANDERSON *et al.*, ‘Attractor states and infrared scaling in de Sitter space’, *Phys. Rev. D* **62**, 124019, 2000; F. FINELLI *et al.*, ‘Adiabatic regularization of the graviton stress-energy tensor in de Sitter space-time’, *Phys. Rev. D* **71**, 023522, 2005; L. R. ABRAMO, R. BRANDENBERGER and V. F. MUKHANOV, ‘Energy momentum tensor for cosmological perturbations’, *Phys. Rev. D* **56**, 3248, 1997.
- [22] S. W. HAWKING, ‘Black-hole explosions?’, *Nature* **248**, 30, 1974.
- [23] T. C. BUNCH and P. C. W. DAVIES, ‘Quantum field theory in de Sitter space - Renormalization by point splitting’, *Proc. R. Soc. London A* **360**, 117, 1978.
- [24] N. N. BOGORIUBOV, ‘A new method in the theory of superconductivity III’, *Sov. Phys. JETP* **7**, 51, 1958.
- [25] A. VILENKIN and L. H. FORD, ‘Gravitational effects upon cosmological phase transitions’, *Phys. Rev. D* **26**, 1231, 1982.
- [26] C. BERNARD and A. DUNCAN, ‘Regularization and renormalization of quantum field theory in curved space-time’, *Ann. Phys. (NY)* **107**, 201, 1977.
- [27] S. COLEMAN and J. MANDULA, ‘All Possible Symmetries of the S Matrix’, *Phys. Rev.* **159**, 1251, 1967.
- [28] Y. A. GOLFANG and E. P. LIKHTMAN, ‘Extension of the algebra of Poincaré group generators and violation of p invariance’, *Sov. Phys. JETP Lett.* **13**, 323, 1971.
- [29] R. G. HAGG, J. T. LOPUSZANSKI and M. F. SOHNIS, ‘All possible generators of supersymmetries of the S-matrix’, *Nucl. Phys. B* **88**, 257, 1975.
- [30] M. A. PESKIN and D. V. SCHROEDER, *An introduction to quantum field theory*, Addison-Wesley, 1995.
- [31] D. BAILIN and A. LOVE, *Supersymmetric gauge field theory and string theory*, Institute of Physics Publishing, 1996.
- [32] F. COOPER, A. KHARE and U. SUKHATME, ‘Supersymmetry and quantum mechanics’, *Phys. Rep.* **251**, 267, 1995.
- [33] J. WESS and B. ZUMINO, ‘A Lagrangian model invariant under supergauge transformations’, *Phys. Lett. B* **49**, 52, 1974; ‘Supergauge transformations in four dimensions’, *Nucl. Phys. B* **70**, 39, 1974.
- [34] A. BRIGNOLE, L. E. IBAÑEZ and C. MUÑOZ, ‘Soft supersymmetry-breaking terms from supergravity and superstring models’, in ‘Perspectives on Supersymmetry’, G. Kane (ed.), World Scientific, 1998.
- [35] L. O’RAIFEARTAIGH, ‘Spontaneous symmetry breaking for chirals scalar superfields’, *Nucl. Phys. B* **96**, 331, 1975.
- [36] P. FAYET and J. ILIOPoulos, ‘Spontaneously broken supergauge symmetries and goldstone spinors’, *Phys. Lett. B* **51**, 461, 1974.
- [37] P. FAYET, ‘Supersymmetry and weak, electromagnetic and strong interactions’,

- Phys. Lett. B* **64**, 159, 1976; ‘Spontaneously broken supersymmetric theories of weak, electromagnetic and strong interactions’, *Phys. Lett. B* **69**, 489, 1977; ‘Relations between the masses of the superpartners of leptons and quarks, the goldstino coupling and the neutral currents’, *Phys. Lett. B* **84**, 416, 1979.
- [38] P. FAYET, ‘Supersymmetric theories of particles and interactions’, *Phys. Rep.* **105**, 21, 1984.
- [39] K. A. OLIVE, *Supersymmetry*, in *The primordial Universe – Les Houches Session LXXI*, Editors P. Binétruy *et al.*, EDP Science & Springer, 2000.
- [40] P. van NIEUWENHUIZEN, ‘Supergravity’, *Phys. Rep.* **68**, 189, 1981.

11

Phase transitions and topological defects

From a cosmological perspective, one of the most important implications of grand unified theories is the symmetry-breaking scheme, that is the precise way in which the internal symmetries of particle-physics theories evolve dynamically towards the symmetries of the electroweak standard model. These symmetry breakings occur via phase transitions¹ that potentially lead to the formation of topological defects, whose numerous cosmological consequences can in turn indirectly impose constraints on the way the grand unification could be implemented [2].

11.1 Phase transitions

In the early phases of the Universe, when the density was very large, the temperature may have reached, or even exceeded, the grand unification scale. If interactions are described by a unified theory then the symmetry was initially not broken. This symmetry qualitatively changes below the critical temperature and so we have a *phase transition*.

To understand the influence of temperature on the Higgs field potential, we should compute the effective potential for this field, taking into account its interactions with the particles in the environment, in a way very similar to the computation of the running of the coupling constants with energy. In the case at hand, the temperature plays the role of the energy so that the effective potential depends on the temperature. This is why thermal effects come into play in the context of field theory and how they can be computed.

A complete treatment of quantum field theory at finite temperature is beyond the scope of this book, and can be found in Ref. [3], for example. For completeness and self-consistency, however, we summarize here the main points necessary for our purposes.

11.1.1 Thermal field theory

11.1.1.1 Symmetry breaking and the classical part of the field

To compute the effects of high temperatures, we should start by describing physics at zero temperature. For this, we choose a model in which the Higgs field, ϕ , is in an arbitrary representation of the symmetry group \mathcal{G} . It therefore has an arbitrary

¹A detailed discussion of phase transitions in non-cosmological contexts, but where concepts such as scale invariance are considered, can be found in Ref. [1].

number of degrees of freedom and, at zero temperature, evolves with a self-interacting potential of the form

$$V(\phi) \Big|_{T=0} = \lambda \left(|\phi|^2 - \eta^2 \right)^2, \quad (11.1)$$

where $|\phi|^2 \equiv \phi^\dagger \phi$. The vacuum state of this theory should be invariant under all possible transformations of the Poincaré group and so cannot depend on any space-time coordinates. The vacuum corresponds to a minimum of the potential, which imposes a field vacuum expectation value (VEV)

$$\frac{dV}{d\phi} = 0 \implies \phi = \phi_0, \quad \text{with} \quad |\phi_0|^2 = \eta^2, \quad (11.2)$$

where ϕ_0 represents a fixed constant configuration. In the case for which the theory is simply Z_2 invariant, and thus $\phi \in \mathbb{R}$, there are only two possibilities, namely $\phi_0 = \pm \eta$. For a broken U(1) invariance with $\phi \in \mathbb{C}$, we obtain $\phi_0 = \eta e^{i\alpha}$, where $\alpha \in \mathbb{R}$ is an arbitrary phase. The symmetry is indeed broken since the vacuum solution is no longer invariant under the transformations of \mathcal{G} .

Let us emphasize that ϕ_0 is a number, or a set of numbers, and not an operator: ϕ_0 is interpreted as the expectation value of the Higgs field in the vacuum $|0\rangle$, i.e. $\phi_0 = \langle 0 | \phi | 0 \rangle$. We can thus decompose the field as

$$\phi = \phi_c + \hat{\phi}, \quad (11.3)$$

where $\phi_c = \phi_0$ is a number, or a function of space-time coordinates, the so-called ‘classical’ part, and $\hat{\phi}$, which is its ‘quantum’ part, is an operator.

11.1.1.2 Finite temperature potential

The Lagrangian of a general theory contains scalar bosons φ_i , vector bosons, A_μ^a and fermions, ψ^r , all of them coupled to the Higgs field as in the case of the standard model described in Chapter 2. These couplings generally contain a quadratic part, thereby leading to terms interpretable as mass terms for all the particles. These masses therefore depend on the classical part of the Higgs field ϕ_c , and we have, in general,

$$\mathcal{L}^{\text{quad}} = -\frac{1}{2} [M_s^2(\phi_c)]_{ij} \varphi_i \varphi_j - \frac{1}{2} [M_v^2(\phi_c)]_{ab} A^{a\mu} A_\mu^b - [M_f(\phi_c)]_{rs} \bar{\psi}^r \psi^s, \quad (11.4)$$

where the mass matrices M_s , M_v and M_f are determined by the representations in which the different particles are initially placed.

Given the theory at zero temperature, it is now possible to write down the thermal corrections and determine the effective potential at high temperature, under the approximation in which all the particles in the thermal bath are considered as ultra-relativistic. We find (see Ref. [3] for technical details)

$$V_{\text{eff}}(\phi_c, T) = V(\phi_c) \Big|_{T=0} + \Delta V(\phi_c, T),$$

where $V(\phi_c) \Big|_{T=0}$ is given by equation (11.1), and

$$\Delta V(\phi_c, T) = -\frac{\pi^2 T^4}{90} \left(N_B + \frac{7}{8} N_F \right) + \frac{T^2}{24} [\text{Tr}(M_s^2) + 3\text{Tr}(M_v^2) + 2\text{Tr}(M_F^2)], \quad (11.5)$$

is the thermal correction. The numbers N_B and N_F represent, respectively, the bosonic and fermionic degrees of freedom ($N_B = 1$ for a scalar field, $N_B = 2$ for a massless gauge boson, $N_B = 3$ for a massive gauge boson, and $N_F = 4$ for a Dirac spinor).

11.1.2 Dynamics of the symmetry breaking

11.1.2.1 Example of the Abelian Higgs model

The mechanism of symmetry breaking can be illustrated by considering the simplest model of symmetry breaking, namely that of a U(1) invariance, broken by a complex Higgs field ϕ

$$\mathcal{L}_{\text{ab.H.}} = -(D_\mu \phi)^* D^\mu \phi - \frac{1}{4} H_{\mu\nu} H^{\mu\nu} - V(\phi) \Big|_{T=0}, \quad (11.6)$$

where the covariant derivative is $D_\mu \phi = (\partial_\mu + iqB_\mu)\phi$ and $H_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$. In this simplified model, we introduce no fermions, although they can be added without qualitatively changing the results. Decomposing the field as²

$$\phi = \phi_c + \frac{1}{\sqrt{2}} (\hat{\varphi}_1 + i\hat{\varphi}_2) = \frac{1}{\sqrt{2}} (\phi_1 + i\phi_2 + \hat{\varphi}_1 + i\hat{\varphi}_2),$$

the quadratic part (11.4) is given by

$$\mathcal{L}_{\text{ab.H.}}^{\text{quad}} = -q^2 |\phi_c|^2 B^\mu B_\mu - \lambda [(\phi_1^2 + |\phi_c|^2 - \eta^2) \hat{\varphi}_1^2 + (\phi_2^2 + |\phi_c|^2 - \eta^2) \hat{\varphi}_2^2], \quad (11.7)$$

where we have used $|\phi_c|^2 = \frac{1}{2} (\phi_1^2 + \phi_2^2)$. We easily read from this relation between the mass matrices and their traces, namely

$$\text{Tr}[M_s^2(\phi_c)] = 4\lambda (2|\phi_c|^2 - \eta^2),$$

and

$$\text{Tr}[M_v^2(\phi_c)] = 2q^2 |\phi_c|^2,$$

from which we now obtain the dynamics of the transition.

11.1.2.2 Critical temperature

Having obtained these results, we deduce the following thermal correction to the scalar field effective potential

$$V_{\text{ab.H.}}^{\text{eff}} = \lambda [| \phi_c |^2 - \bar{\eta}^2(T)]^2 + f(T, \eta), \quad (11.8)$$

where the function f is given by

²We have chosen directly here the classical part to be real: this merely simplifies the computations while leaving unchanged the results, since the underlying theory is invariant under U(1) transformations of the Higgs field phase.

$$f(T, \eta) = \frac{T^2 \eta^2}{12} (3q^2 - 2\lambda) - \frac{T^4}{9} \left[\frac{\pi^2}{9} + \frac{(4\lambda + 3q^2)^2}{64\lambda} \right].$$

This function only depends on temperature (not on ϕ_c), and so does the minimum $\bar{\eta}$, through the relation

$$\bar{\eta}^2(T) = \eta^2 - \frac{T^2}{24\lambda} (4\lambda + 3q^2), \quad (11.9)$$

which, as it turns out, is not positive definite. It actually vanishes when the temperature decreases to its critical value T_c , given by

$$T_c^2 = \frac{24\lambda\eta^2}{4\lambda + 3q^2}. \quad (11.10)$$

Above this temperature, the potential has a unique global minimum for $\phi_c = 0$, while it is shifted to a set of non-vanishing values of ϕ_c for $T < T_c$. Figure 11.1 illustrates the temperature dependence of the potential.

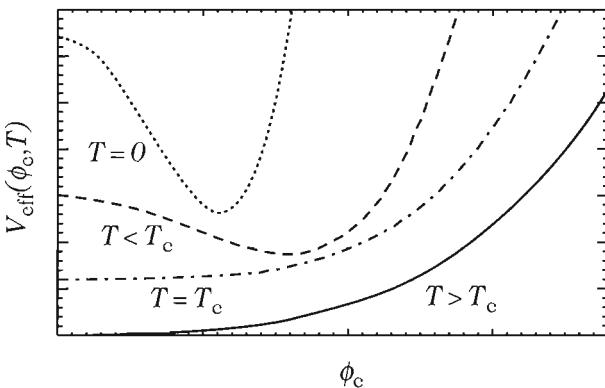


Fig. 11.1 Variation of the potential $V_{\text{eff}}(\phi_c, T)$ as a function of the ‘classical’ part of the scalar field ϕ_c for different temperatures T . We have only represented a single variable for the scalar field, which is actually one of its components, i.e., the imaginary or real part (or an arbitrary combination of them) in the case of a complex field. At high temperatures ($T > T_c$), the potential has a unique minimum at the symmetry-restoring point $\phi_c = 0$ that evolves, as the temperature decreases, to become a maximum. A minimum eventually develops, and the potential subsequently smoothly transforms into the zero-temperature curve (dotted line). Since the temperature of the Universe decreases during the expansion, the symmetry naturally breaks down, hence spontaneously producing a phase transition.

11.1.2.3 Effective potential

At high temperatures, i.e. for $T > T_c$, the effective potential has a unique minimum localized at $\phi_c = 0$ (Fig. 11.1). The symmetry is thus unbroken as long as the temperature is sufficiently high. When the temperature decreases due to the effect of another

physical mechanism, such as, for instance, the expansion of the Universe, the shape of the potential changes until its minimum is no longer at $\phi_c = 0$. The symmetry is then spontaneously broken (note that in the cosmological framework, and actually only in this case, the term spontaneous, suggesting some dynamics, is completely appropriate). This is a real phase transition, in the thermodynamical sense, since the order parameter, here the value of the field ϕ_c , only varies with the temperature until the critical temperature is reached. The potential (11.8) is then minimized for $\phi_c = \bar{\eta}(T) = \eta\sqrt{1 - (T/T_c)^2}$ as long as $T \leq T_c$.

The form of this potential, in particular its temperature dependence, can be understood in terms of thermal theory once fluctuations are taken into account. When the temperature is low, the field must choose one of the possible minima, and its energy is not sufficient to cross over the potential barrier. Such a transition is only allowed by tunnel effects that are all the more improbable the higher the barrier. At high temperatures, the interaction of the field with the thermal bath provides an additional energy, coming from the particle bath. If the energy is greater than a given value, known as the Hagedorn temperature, then the thermal fluctuations allow the field to cross over the potential barrier. The probabilities that the field remains in this state or returns to its initial state are then identical, so it oscillates between the two states with no preferred one. This is all the more true the higher the temperature since the transition rate increases with the temperature.

From the classical part of the field $\phi_c(\mathbf{x}, t)$, one can define an expectation value $\bar{\phi}_c \equiv \langle \phi_c(\mathbf{x}, t) \rangle$, averaged over times or spatial scales large compared to the microscopic parameters (the transition time or the correlation length). This is what we call a ‘coarse-grained’ description since we take the average over small but finite-size regions. The field $\phi_c(\mathbf{x}, t)$ oscillates between the two vacuum values at zero temperature, so that $\bar{\phi}_c$ must vanish. This translates into an additional mass term in the effective potential, forcing the average field to have a vanishing value.

11.1.2.4 Other corrections

Besides the quadratic part (11.4) of the Lagrangian, thermal corrections can also lead to cubic terms in the scalar field. This term depends on the particle content of the theory at zero temperature. We find

$$\Delta V^{(3)}(\phi_c, T) = \beta T |\phi_c|^3, \quad (11.11)$$

where the constant β depends on the initial coupling constants of the theory. The complete effective potential, obtained in the Abelian Higgs model from the sum of the corrections (11.8) and (11.11), is represented in Fig. 11.2 for the two cases $\beta < 0$ and $\beta > 0$. In the first case, $\beta < 0$, the phase transition is of first order, while for $\beta > 0$ it is of second order.³

³A transition is said to be of second order when it happens in a continuous way, the phase changing continuously through the critical point. By contrast, a first-order transition is characterized by the fact that the different phases can coexist at the critical point. According to the Ehrenfest classification, a transition is of order n if the n th derivative of an appropriate thermodynamic potential (often the free energy) is discontinuous, while all the derivatives of order $p < n$ are continuous when crossing the transition point.

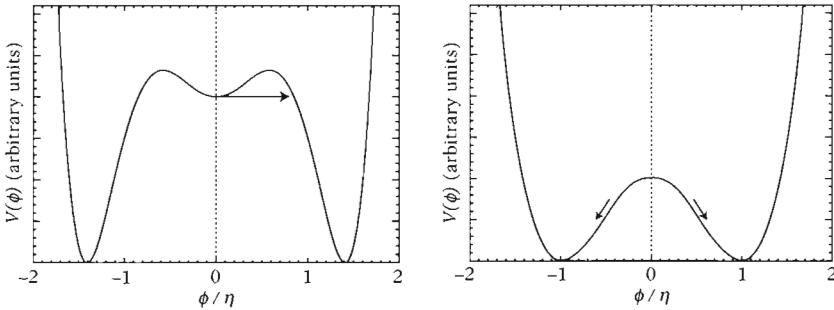


Fig. 11.2 (left): Potential $V(\phi)$ for a scalar field ϕ leading to a symmetry breaking via a first-order phase transition. The phase $\phi = 0$ is metastable, so that the transition must occur via a tunnel effect. The field, which is initially trapped at the central value, in the local minimum, cannot cross over the potential barrier classically, but quantum fluctuations can allow it to cross at once by tunnelling (along the arrow indicated in the graph, for instance). The phase transition occurs through nucleation. (right): Potential $V(\phi)$ for a scalar field ϕ leading to a symmetry breaking via a second-order phase transition. The phase $\phi = 0$ is unstable. As soon as the temperature becomes lower than the critical value, the field must choose a direction to break the symmetry, following one of the two arrows indicated in the graph.

11.1.3 Formation of topological defects

From now on, we will focus on the classical part, ϕ_c , of the field. To avoid unnecessarily heavy notations, we simply denote it by ϕ , bearing in mind that the fields in question are now classical.

11.1.3.1 General symmetry-breaking scheme

The symmetry-breaking mechanism based on a scalar field will always be presented in what follows by a scheme

$$\boxed{\mathcal{G} \xrightarrow{\phi_0} \mathcal{H} \implies \mathcal{M} \sim \mathcal{G}/\mathcal{H}}, \quad (11.12)$$

in which the initial invariance under transformations of the group \mathcal{G} is reduced, due to the appearance of a non-vanishing vacuum expectation value for the Higgs field in the representation ϕ_0 , to that of the subgroup \mathcal{H} , hence defining the vacuum manifold \mathcal{M} as being isomorphic to the quotient group \mathcal{G}/\mathcal{H} (see Section 2.5.2 of Chapter 2 for some examples).

11.1.3.2 Quotient group

Before explicitly focusing on the vacuum topology, let us quickly recall some useful definitions. More details can be obtained in Ref. [4].

Let us begin with the notion of an equivalence coset of an equivalence relation \mathcal{R} (defined as being reflexive, symmetric and transitive). Defining the equivalence relation $g \mathcal{R} g'$ between two elements g and g' of \mathcal{G} by

$$g \mathcal{R} g' \iff \exists h \in \mathcal{H}; g' = gh,$$

the equivalence coset $[g]$ of this relation is $[g] = \{gh; h \in \mathcal{H}\}$. We also denote it by $g\mathcal{H}$, called a ‘left coset’. In the same manner we can define a ‘right coset’ $\mathcal{H}g$. It is easy to show that for two elements g and g' of \mathcal{G} , we have either $g\mathcal{H} \cap g'\mathcal{H} = \emptyset$ or $g\mathcal{H} = g'\mathcal{H}$, so that the set consisting of all $g\mathcal{H}$ forms a decomposition of \mathcal{G} on \mathcal{H} into disjoint equivalence cosets. This set is the quotient space.

If $ghg^{-1} \in \mathcal{H} \forall g \in \mathcal{G}$ and $\forall h \in \mathcal{H}$, then the subgroup \mathcal{H} of \mathcal{G} is said to be ‘normal’ (or ‘invariant’), which is denoted by $\mathcal{H} \triangleright \mathcal{G}$. In this particular case, the left and right cosets are the same, and there is no need for any distinction.

If $\mathcal{H} \triangleright \mathcal{G}$, then the quotient space \mathcal{G}/\mathcal{H} is a group, called the quotient group, whose group operation denoted by $*$, is defined by $(g\mathcal{H}) * (g'\mathcal{H}) = (gg')\mathcal{H}$.

11.1.3.3 Topological defects

This structure implies the possibility of the existence of topological defects, which are field configurations for which there exist domains where the symmetry is left unbroken, i.e. for which $\phi = 0$. They are localized regions of space for topological reasons.

Since such a configuration does not correspond to a minimum of the potential, it contains a localized energy density around the points for which $\phi = 0$. Although we will discuss this in more detail later, we can immediately remark that the condition $\phi = 0$, imposed in a three-dimensional space, can allow for defects of different dimensions. For instance,

- if $\phi \in \mathbb{R}$, then $\phi(\mathbf{x}) = 0$ is a relation between three coordinates and defines a surface, called a ‘domain wall’, which is unfortunately an inappropriate terminology since it suggests a kind of rigidity that these walls do not have;⁴
- if $\phi \in \mathbb{C}$, then this equation reduces to $\Re[\phi(\mathbf{x})] = 0$ and $\Im[\phi(\mathbf{x})] = 0$. We thus have two equations for three coordinates, and therefore the solution of the equation $\phi = 0$ is at the intersection of the two membranes: it is a 1-dimensional object, called a vortex line, by analogy with the defects appearing in similar systems of condensed-matter physics, or also cosmic strings, which is the terminology we will keep;
- if ϕ is a vector in $\text{SO}(3)$, for example, with three components, then $\phi = 0$ gives three relations for three coordinates. There is only one solution, defining a point, called a monopole.

In a space with more than three dimensions, such as encountered in particular in string theory, higher-dimensional defects can be thought of, having more dimensions. In particular, it has even been suggested that our own 3-dimensional space could be such a topological defect, for instance, a domain wall in a 5-dimensional space [5], or a string in 6 dimensions [6] (see Chapter 13).

⁴The term membrane, suggesting a greater degree of flexibility, seems more adapted to the description of the dynamics of these objects. We shall, however, use the usual terminology in what follows.

11.2 Domain walls

Let us start with the simplest example of a real scalar field, with a Lagrangian invariant under transformations of the group $\mathcal{G} = Z_2 = \{-1, +1\}$. This condition amounts to imposing that the theory is invariant under the transformation $\phi \rightarrow -\phi$. At zero temperature, we thus consider the Lagrangian

$$\mathcal{L}_{\text{walls}} = -\frac{1}{2}\partial_\mu\phi\partial^\mu\phi - \lambda(\phi^2 - \eta^2)^2. \quad (11.13)$$

The lowest-energy configuration is reached for $\phi = \pm\eta$, the sign choice being irrelevant for any calculation of local physics (cross-sections, for instance). This choice is, however, essential to make any computation since it defines the vacuum in a complete way.

No matter what the chosen vacuum value, this value completely breaks the initial symmetry since it is no longer possible to arbitrarily change the field value without changing the configuration. This means that the invariance \mathcal{H} resulting from this breaking is simply the set containing the identity alone, $\mathcal{H} = \{\text{Id}\}$. The breaking scheme equivalent to (11.12) is thus $Z_2 \rightarrow \{\text{Id}\}$. The vacuum manifold is thus $\mathcal{M} \sim Z_2$, which amounts to saying that each element of Z_2 is a possible candidate for describing the vacuum configuration.

11.2.1 Correlation length

Let us now consider how the phase transition occurs while the Universe cools down. Initially, the temperature is very high, so that the scalar field vanishes: by definition, this is the symmetric phase. As the temperature decreases and the potential becomes deeper, the field tends to follow the inclination of the potential, and thus to take a non-zero value. Thus, it must choose in which direction, i.e. along $\phi > 0$ or $\phi < 0$. This choice is of course random. In particular, as the field can only interact with itself over finite distances, because of causality, there must exist a *correlation length*, denoted by ξ , beyond which, by definition, the choices of the Higgs fields are uncorrelated. This correlation length is computed as a function of the order of the symmetry-breaking phase transition.

11.2.1.1 First-order transitions

A first-order transition is performed via *nucleation* of the new phase within the old one. To achieve this, the effective potential should only have a local minimum at $\phi = 0$, which is the case if $\beta < 0$ in the thermal correction (11.11). The field is initially localized in its global minimum, $\phi = 0$. This point becomes a local minimum below the critical temperature (Fig. 11.2). Classically, the field cannot cross over the potential barrier, and it thus remains trapped for some time in the state $\phi = 0$. However, this state is metastable, and quantum fluctuations eventually allow the field to tunnel from one side to the other. This is what the arrow represents in Fig. 11.2, where the choice has been made towards positive values.

If, at a given point in space, the fluctuations are sufficient for the tunnelling to happen, then the field crosses the barrier, say towards positive values, to quickly reach the minimum $\phi = +\eta$. Due to the self-interaction of the field around this point, and

since $\phi = 0$ is not stable, the probability to pull this field towards this minimum increases, so that the region in which $\phi = +\eta$ grows. A bubble of ‘true vacuum’ is said to nucleate spontaneously in the ‘false vacuum’ space. For this reason, this mechanism is called nucleation. Since the energy contained in the true vacuum is, by definition, lower than that contained in the false vacuum, the bubbles grow rapidly, collide, and eventually occupy the entire available space (all the Universe that concerns us in the cosmological context). The phase transition is then complete.

11.2.1.2 Nucleation

The nucleation rate per unit of time and space, $\gamma(T)$, depends of course on the temperature as well as the probability of the field configurations. We find

$$\gamma(T) = A(T)e^{-S_E(T)},$$

where, for dimensional reasons, $A(T) \propto T^4$. The quantity S_E is the Euclidean action, i.e. the action computed by replacing the time t by the purely imaginary variable $t_E = it$. This operation replaces the Minkowski metric by a completely Euclidean metric $ds_E^2 = dt_E^2 + d\mathbf{x}^2$. This transformation is very useful to compute, in particular, the probability of a transition via a tunnel effect.

The correlation length ξ is then simply the characteristic distance between two nucleation sites, and, in a given volume V , the number of nucleated bubbles is of order $V\xi^{-3}$. If the expansion speed of each bubble is of the order of the speed of light, then the correlation length depends on the only parameter at hand, namely, γ . We then find $\xi \propto \gamma^{-1/4}$, evaluated for the phase transition temperature. This is how it is possible to describe a first-order transition only from the knowledge of the microscopic Lagrangian, and from this to make macroscopic observational predictions, which are themselves very dependent on the correlation length.

11.2.1.3 Second-order transition

A second-order phase transition is characterized by the fact that, unlike a first-order transition, it is realized in a continuous way. It is not necessary here to rely on the tunnel effect, and the transition starts classically. Nothing seems to indicate the existence of a preferred length in such a case, so that the correlation length seems infinite. How can we then define it in a consistent way?

The cosmological argument, initially proposed in 1976 by Kibble, amounts to observing that from the initial singularity at $t = 0$ until the transition time, there has been in principle a finite time, so that there must exist a particle horizon (see Section 3.5 of Chapter 3). As long as the dynamics of the Universe is governed by a power law scale factor, $a \propto t^\alpha$, this horizon is of the order of the Hubble radius [see (3.107)]. Causality thus fixes an upper bound for ξ .

This argument, which is based on causality, does not apply to the phase transitions that happen in almost mathematically equivalent laboratory systems, such as liquid crystals or the transition from the normal state to the superfluid state during the cooling of liquid helium. As a matter of fact, we actually observe a finite correlation length in these systems when performing the transition. The system spontaneously

defines a length, ξ_v , which will be the ‘true’ correlation length, from which macroscopic consequences will follow, long after the transition.

11.2.1.4 Beyond causality: the ‘quench’ time

A crucial remark has been made by Kibble and Zurek: the transition usually lasts for a finite time. Moreover, in most physical systems, an equivalent of the particle horizon also exists. It is the sound horizon, defined by the fact that no information, in a given system, can propagate faster than the propagation speed of the perturbations of this system, which only depends on the microscopic physics. This speed is usually of the order of the speed of sound (or of the second-sound speed in the case of helium). The speed at which the phase transition happens is not related to this microscopical characteristic in itself, since it is either imposed from outside (speed at which the experimenter decides arbitrarily to lower the temperature), or given by other physics, for instance, the expansion of the Universe in the cosmological case.

Apart from the phase transition, the correlation length is defined as the distance scale of the thermal fluctuations beyond which two field configurations are equiprobable. At very high temperatures, the field can fluctuate very easily and ξ is thus very small. In the neighbourhood of the transition, at a temperature denoted by T_c , ξ becomes infinite.

Taylor-expanding the bath temperature, $T(t)$, at first order, we define the parameter

$$\epsilon \equiv 1 - \frac{T}{T_c} = \frac{t - t_c}{\tau_Q},$$

where t_c is the time of the transition and τ_Q , which is called the ‘quench’ time, represents the characteristic duration of the transition. Statistical physics [7] then teaches us that ξ diverges as a power law

$$\xi = \xi_0 |\epsilon|^{-\nu},$$

where ν is a ‘critical exponent’. Close to the transition, the system freezes more and more, i.e. the time it needs to reach equilibrium, in other words its relaxation time, τ , increases. It thus also behaves as a power law

$$\tau = \tau_0 |\epsilon|^{-\mu},$$

defining a second critical coefficient μ . Note here that ξ is the correlation length at equilibrium, i.e. for a system that always has the time to relax, in other words for a system whose relaxation time is always small compared to the characteristic time of the transition. This condition cannot be maintained when approaching the transition since then $\tau \rightarrow \infty$. In practice, at the transition, the system cannot maintain equilibrium.

In these conditions, the maximal speed at which information can propagate is

$$c(T) = \frac{\xi(T)}{\tau(T)} = \frac{\xi_0}{\tau_0} |\epsilon|^{\mu-\nu}.$$

For cosmological transitions, this speed is of the order of the speed of light but can be very different for other systems. The sound horizon, like the event horizon in general relativity, is a global quantity defined by

$$h(t) \equiv \int^t c[T(t')] dt' = \frac{1}{1 + \mu - \nu} \frac{\xi_0 \tau_Q}{\tau_0} |\epsilon|^{1+\mu-\nu}.$$

It allows us to obtain the real correlation length by identifying, $\xi_v = h$,

$$\xi_v = \xi_0 \left[(1 + \mu - \nu) \left(\frac{\tau_0}{\tau_Q} \right) \right]^{-\frac{\nu}{1+\mu}} \propto \left(\frac{\tau_Q}{\tau_0} \right)^{\frac{\nu}{1+\mu}}. \quad (11.14)$$

Just like the case of a first-order transition, the correlation length is determined by quantities describing the microscopic physics (the critical exponents are in theory computable), as well as by the transition speed. This relation can be verified experimentally. This has actually been done to a very good accuracy in numerous systems: the density of topological defects produced during such a transition depends explicitly on ξ_v , and this density can be measured.

Whether the transition is of first or second order, we see that it is necessary to know all the microphysical details to compute the correlation length. At the level of approximation required to make any valid prediction in cosmology, we note that it is often sufficient to limit ourselves to

$\xi \simeq T_c^{-1}.$

(11.15)

To a first approximation, we thus associate to the phase transition a corresponding length determined by the critical temperature. From a dimensional point of view, this is the first quantity we can think of.

11.2.1.5 Kibble mechanism for domain walls

Let us consider the system defined by the Lagrangian (11.13) describing a domain wall. Once the transition happens, the scalar field ϕ randomly takes the value η or $-\eta$ at each point. Long after the transition, thermal fluctuations have become negligible. Two regions located at a distance $D > \xi$ apart (where $\xi = \xi_v$ if it is a second-order transition) are then, by definition, uncorrelated. This means that the value of the field in one region can in no case depend on the value in the other. In other words, there is a probability of $\frac{1}{2}$ in each region of volume ξ^3 for the field to choose one value or another, i.e. given two arbitrary regions, a probability of $\frac{1}{2}$ that the values chosen by the field are different.

When two regions where the field is different are put in contact, whether it is when two expanding bubbles meet for a first-order transition or during the dynamics of the second-order transition, the field, which is continuous, must interpolate between the two vacua (Fig. 11.3). In other words, the field cannot remain in its true vacuum everywhere and the true vacuum regions ($\phi = \pm\eta$) are separated by interfaces where $\phi = 0$: domain walls have formed.

11.2.2 Static configurations

Let us now look for the static configurations of the theory (11.13). The Euler–Lagrange equation then reduces to a Klein–Gordon equation

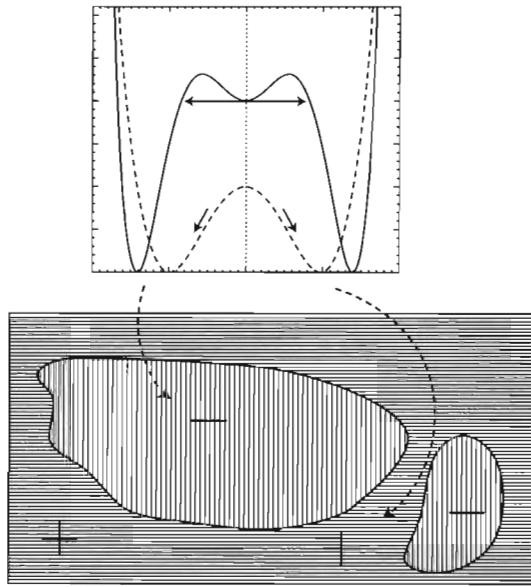


Fig. 11.3 Appearance of regions with different Higgs field values. Independently of the order of the transition, once it is completed, regions with $\phi > 0$ and others with $\phi < 0$ statistically fill the same volume. The distance separating these different regions is, statistically, of the order of the correlation length at the time of the phase transition. The separation surfaces, in the case represented here where the symmetry is discrete, extend in the third dimension and are domain walls, i.e. 2-dimensional.

$$\square\phi = \frac{dV(\phi)}{d\phi} \quad \Rightarrow (-\partial_t^2 + \nabla^2)\phi = 4\lambda\phi(\phi^2 - \eta^2), \quad (11.16)$$

when neglecting the expansion of the Universe. For a static configuration, ϕ is independent of t and we recover the simplest solution $\phi = \pm\eta$. Moreover, we note that $\phi = 0$ is still a solution (it is the local maximum of the potential). In any case, the solution is also space independent.

Let us now look for a domain-wall solution since we have seen that this theory produces some during the phase transition. Very close to the wall, we can consider it as being planar. Imposing a translation symmetry in the (x, y) plane, the Klein-Gordon equation (11.16) then reduces to

$$\frac{d^2\phi}{dz^2} = 4\lambda\phi(\phi^2 - \eta^2), \quad (11.17)$$

for static configurations. The non-trivial solution to this equation is

$$\phi(z) = \eta \tanh\left(\frac{z}{z_c}\right), \quad \text{where} \quad z_c \equiv \left(\sqrt{2\lambda}\eta\right)^{-1}, \quad (11.18)$$

which interpolates asymptotically, i.e. for $z \rightarrow \pm\infty$, between the two possible true vacua, $\phi = \pm\eta$. Using (1.107) or (2.78), we can compute the energy density profile of this field configuration,

$$\rho_{\text{wall}} = \frac{1}{2} \left(\frac{d\phi}{dz} \right)^2 + V(\phi) = \frac{2\lambda\eta^4}{\text{ch}^4(\sqrt{2\lambda}\eta z)}. \quad (11.19)$$

The solution (11.18), as well as the energy density (11.19) it contains, are represented in Fig. 11.4. The characteristic width of the wall, z_c , is roughly the Compton wavelength associated to the Higgs field responsible for the phase transition. Finally, the energy per unit area of the wall is obtained by integration over the transverse direction z ,

$$U_{\text{wall}} \equiv \int_{-\infty}^{+\infty} \rho_{\text{wall}}(z) dz = \frac{\eta^3}{\sqrt{2\lambda}} \int_{-\infty}^{+\infty} \frac{dx}{\text{ch}^4 x} = \frac{2}{3} \sqrt{\frac{2}{\lambda}} \eta^3. \quad (11.20)$$

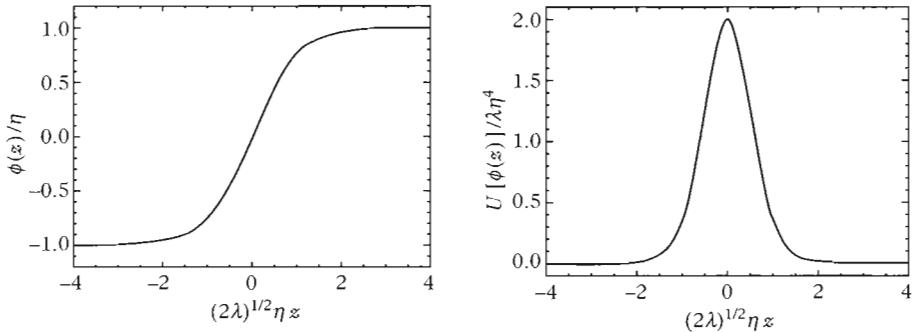


Fig. 11.4 (left): Exact solution interpolating between the two possible vacuum states of the Higgs field for the Z_2 symmetry breaking. This is a domain wall aligned in the (x, y) plane, so that the field only depends on the distance to the wall, i.e. z . (right): Energy density contained in the domain-wall configuration as a function of the distance z to the centre of the wall.

This relation first shows that the defect is clearly a non-perturbative object since its total energy is inversely proportional to the square root of the coupling constant λ , and thus no Taylor expansion around $\lambda = 0$ can lead to it. It also provides the order of magnitude of the energy density as the cube of the symmetry breaking scale. For a phase transition at the grand unification scale, with $\eta \sim 10^{15} \text{ GeV}$, this leads to a value $U_{\text{wall}} \sim 4.6 \times 10^{49} \text{ kg} \cdot \text{m}^{-2}$, that is around $10^{13} M_\odot$ per square millimeter, when converting to standard units!

11.3 Vortices

Let us now consider the case of an initial $U(1)$ symmetry with breaking scheme given by $\mathcal{G} = U(1) \mapsto \{\text{Id}\} = \mathcal{H}$. This scheme corresponds to the case already studied of a complex scalar field in a Higgs potential, i.e. the Abelian Higgs model (11.6).

11.3.1 Kibble mechanism for cosmic strings

The case of the formation of a vortex, or string, i.e. a 1-dimensional defect containing energy concentrated in a thin tube, is very similar to that of the formation of a wall. The situation is represented in Fig. 11.5, illustrating the case of a second-order phase transition. At different points in space, separated by a distance larger than the correlation length, or equivalently, in different bubbles in the nucleation case, the Higgs field falls in a minimum of the potential, in this case a fixed amplitude and an arbitrary phase.

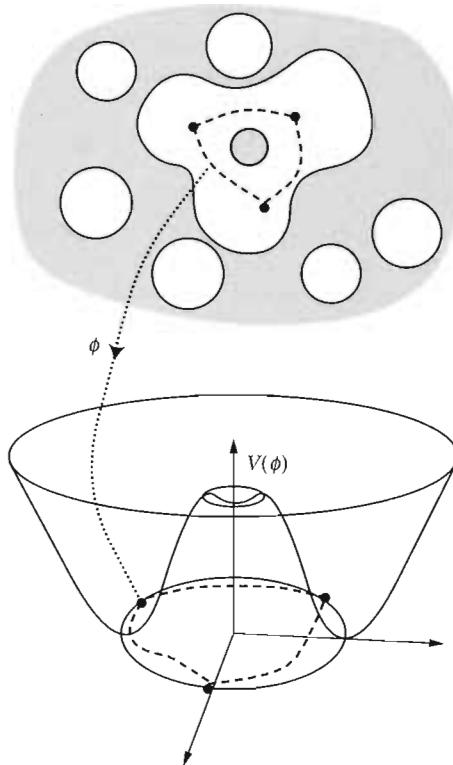


Fig. 11.5 Kibble mechanism for the breaking of the group U(1) and formation of cosmic strings. The upper diagram represents the physical space and the bottom one the field space. After the phase transition, the Higgs field, ϕ , has taken a different value in each point of space. This value is one of the potential minima, $V(\phi)$, which is different in the three regions represented here. If, along a closed path in real space, the Higgs field makes a complete loop in field space, then there must exist a point included inside the path for which this phase is not defined. This implies that the field itself vanishes. In this case, the field is stuck at the maximum of the potential and the string configuration we obtain contains energy. From Ref. [8].

These phases are randomly distributed around the circle of minima since no preferred direction exists at the microscopic level. When three bubbles meet (or at the intersection of three uncorrelated regions), these phases can make one (or several) complete loops around the circle of minima in field space. In this case, we can imagine drawing a closed path Γ , represented in dotted lines in Fig. 11.5, along which the phase of the Higgs field varies continuously. At each point x of Γ , the field can only have a unique phase $\vartheta(x)$, such that one can change gauge locally without modifying the local physics. After a complete loop along Γ , the variation of the phase $\Delta\vartheta$ can only be an integer multiple of 2π :

$$\Delta\vartheta = 2\pi n, \quad \text{with} \quad n \in \mathbb{Z}. \quad (11.21)$$

Assuming from now on that $n \neq 0$ since we want to look at a defect solution, we then note that there must necessarily exist a singular point inside the closed contour Γ . Indeed, if we try to deform the loop Γ continuously into a point, we observe a discontinuity: somewhere the phase must abruptly go from ϑ to $\vartheta + \Delta\vartheta$ and once the curve is brought to a point, this phase can then no longer be defined. The only way for this to happen, is to have $\phi = 0$ at this point.

Another way to understand this phenomena is shown in Fig. 11.6, where the phase of the Higgs field is represented by an arrow (it is a direction in the complex plane). Along the path Γ , this arrow can wind once ($n = 1$) or several (e.g., $n \geq 2$ case) times in space, or even none ($n = 0$), as illustrated. In both cases where $n \neq 0$, this implies the presence of a singularity inside the path, a point where the arrow can no longer be represented. This representation clearly illustrates the analogy with vortex lines that can appear during phase transitions leading to a superconducting system. It is actually no coincidence if the Ginzburg–Landau model that describes such a system, is essentially based on the non relativistic version of the Lagrangian (11.6), with a term of the form (11.11).

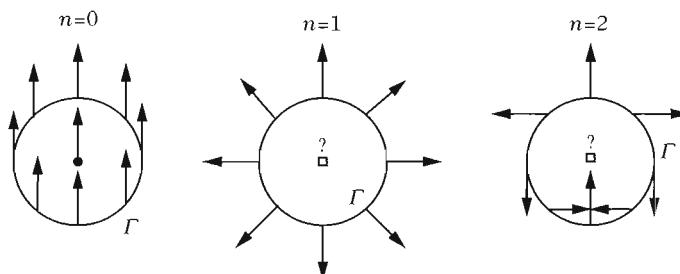


Fig. 11.6 Illustration of the Kibble mechanism for which the phase of the Higgs field is represented by an arrow, equivalent to the magnetic moment in a ferromagnetic system, indicating the direction chosen locally by ϑ . The case $n = 0$ is regular everywhere, and the field is continuous at all points, but both cases $n = 1$ and $n = 2$ necessarily lead to the existence of a singularity somewhere inside the path Γ .

The singular point encountered inside Γ actually extends to form a 1-dimensional line when the graph (Fig. 11.6) is extended to the third dimension of space. The strings

created are then necessarily infinite or in the form of loops. Indeed, let us assume that this is not the case and that a segment of string is formed. It would then be possible to deform Γ in such a way as to make it go around the segment, and then one could freely deform it to finally contract it into a point, in contradiction with the initial hypothesis according to which Γ is precisely not contractible to a point.

In the previous example, the presence of a string is connected with the integer number n obtained from a given loop. This integer, called the winding number or Pontryagin index, arises because the first homotopy group of the $U(1)$ topological space is \mathbb{Z} . It expresses the fact that the vacuum is not simply connected⁵ and shows that strings can be classified along \mathbb{Z} .

11.3.2 Internal structure

The internal structure of a string is obtained in the same way as for a domain wall, up to the fact that there is no known analytic solution to the field differential equations.

Most of the strings discussed in the literature, which are assumed to have a potential role to play in cosmology, are Abelian strings. To describe them, we usually assume that a subgroup $U(1)$ of the grand unification group is broken independently of the rest of the total group. In this case, we neglect most of the degrees of freedom other than those necessary for the description of the string itself, and we end up with an Abelian Higgs model (11.6).

11.3.2.1 Static field equations

Just as for domain walls, we first establish the dynamical equations for the Lagrangian (11.6) for which we decompose as

$$\phi(\mathbf{x}) = \varphi(\mathbf{x})e^{i\alpha(\mathbf{x})}, \quad (11.22)$$

giving

$$\nabla_\mu \nabla^\mu \varphi = (\nabla_\mu \alpha + qB_\mu)(\nabla^\mu \alpha + qB^\mu)\varphi + 2\lambda(\varphi^2 - \eta^2)\varphi, \quad (11.23)$$

for the amplitude of the scalar field,

$$\nabla_\mu [(\nabla^\mu \alpha + qB^\mu)\varphi^2] = 0, \quad (11.24)$$

for the phase, this relation being simply the expression of the conservation of the $U(1)$ current, and

$$\nabla_\mu H^{\mu\nu} = 2q(\nabla^\nu \alpha + qB^\nu)\varphi^2, \quad (11.25)$$

for the gauge field.

We now look for a string configuration for which the phase of the Higgs field winds n times in space. Locally, one can neglect the string curvature and thus consider a coordinate system in which the set of points with $\varphi = 0$ is aligned along the z axis, in

⁵A simply connected space is a space connected by arcs (two arbitrary points can be joined by a path in space) such that any loop (shoe lace) is homotopic to a point.

cylindrical coordinates (r, θ, z) . The string condition is implemented by the Nielsen–Olesen solution [9] for which the phase is simply

$$\alpha = n\theta,$$

and the amplitude only depends, by symmetry, on the radial distance, $\varphi = \varphi(r)$. As for the gauge field, it has only one non-zero component, namely B_θ . Equations (11.23) and (11.25) become [(11.24) is trivially satisfied with this solution],

$$\begin{cases} \frac{d^2 X}{d\rho^2} + \frac{1}{\rho} \frac{dX}{d\rho} = \frac{XQ^2}{\rho^2} + \frac{1}{4} (X^2 - 1) X, \\ \frac{d^2 Q}{d\rho^2} - \frac{1}{\rho} \frac{dQ}{d\rho} = \tilde{q}^2 X^2 Q, \end{cases} \quad (11.26)$$

where in these relations, distances are expressed in units of the Compton wavelength of the Higgs field, defining $\rho \equiv r/r_h = m_h r = \sqrt{8\lambda}\eta r$, and we have defined $X \equiv \varphi/\eta$ and $Q \equiv n + qB_\theta$ in order to deal only with dimensionless variables. These notations as well as the equations they lead to, show that there is only one parameter on which the solution depends, namely $\tilde{q}^2 \equiv q^2/(4\lambda)$.

Equations (11.26) must be solved numerically with boundary conditions corresponding to a vortex line, namely $X(0) = 0$, defining the defect, $X(\infty) = 1$, being the nominal value of the Higgs field in the vacuum, and $Q(0) = n$ giving the Pontryagin index of the string, as well as $dQ(0)/d\rho = 0$ for regularity. One can also impose $Q(\infty) = 0$, which is necessary in order to be in the vacuum far away from the string, leading to the same result. Figure 11.7 illustrates this solution in the case of $n = 1$.

From the solution of these equations, we can obtain the string energy-momentum tensor defined by (1.84), namely

$$\begin{aligned} T_{\mu\nu} = & 2 [\partial_\mu \varphi \partial_\nu \varphi + \varphi^2 (\partial_\mu \alpha + qB_\mu) (\partial_\nu \alpha + qB_\nu)] \\ & - [\partial_\sigma \varphi \partial^\sigma \varphi + \varphi^2 (\partial_\sigma \alpha + qB_\sigma) (\partial^\sigma \alpha + qB^\sigma)] g_{\mu\nu} \\ & + H_{\mu\sigma} H_\nu^\sigma + \left[\lambda (\varphi^2 - \eta^2)^2 - \frac{1}{4} H_{\sigma\beta} H^{\sigma\beta} \right] g_{\mu\nu}, \end{aligned} \quad (11.27)$$

and deduce the integrated quantities that will be necessary for a macroscopic description of the string. Note that it is not useful to compute all the components of the energy-momentum tensor since we know that it is conserved, i.e. $\nabla_\mu T^{\mu\nu} = 0$. Indeed, the r component of this conservation law implies

$$\frac{dT_r^r}{dr} = \frac{T_\theta^\theta - T_r^r}{r},$$

where we have assumed that all the microscopic functions only depend on the radial distance r , which is certainly the case for the Nielsen–Olesen solution. If the components of T_ν^μ decrease with r faster than r^{-2} , which is again the case for a local string as considered here, then the previous relation leads to

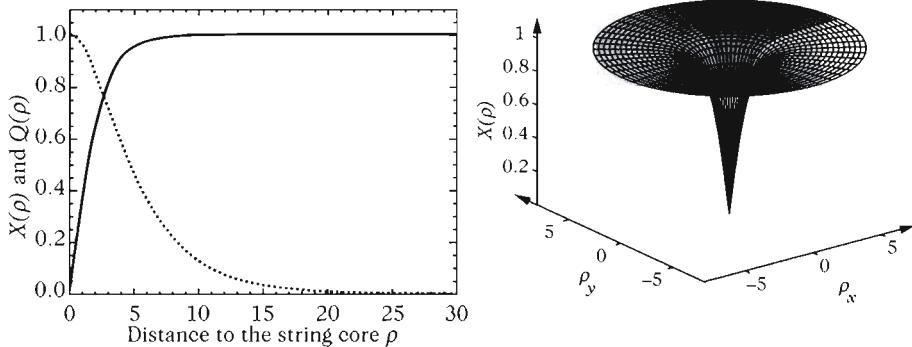


Fig. 11.7 Structure of a cosmic string. (left): Value of the Higgs field and the associated gauge field in the Nielsen–Olesen solution for the parameters $\tilde{q}^2 = 0.1$ and $n = 1$. (right): Representation of the value of the Higgs field amplitude $X(\rho) = \varphi/\eta$ as a function of the distance to the vortex centre, assumed to be at the origin of the coordinate system ($\rho_x = \rho_y = 0$ in the figure). At the centre of the string, where the field vanishes, we have $X(0) = 0$, whereas asymptotically, we recover a vacuum configuration, i.e. $X(\rho_x, \rho_y) \rightarrow 1$ when $(\rho_x, \rho_y) \rightarrow \infty$. Distances ρ , ρ_x and ρ_y are expressed in units of the Higgs-field Compton wavelength $r_h^{-1} = \sqrt{8\lambda\eta}$. The figure on the right is adapted from Ref. [8].

$$\int_0^\infty \frac{dT_r}{dr} r^2 dr = [r^2 T_r]_0^\infty - 2 \int_0^\infty T_r r dr = \int_0^\infty (T_\theta^\theta - T_r^r) r dr,$$

and thus to $\int (T_\theta^\theta + T_r^r) r dr = 0$. Moreover, placing ourselves in Cartesian coordinates (x, y) instead of (r, θ) , rotational invariance around the string imposes that $\int T_x^x r dr = \int T_y^y r dr$. Finally, since $T_x^x + T_y^y = T_\theta^\theta + T_r^r$, we obtain $\int T_x^x r dr = \int T_y^y r dr = 0$. Since the energy-momentum tensor does not depend on the system of coordinates, we conclude that only two integrated components are meaningful. We can choose

$$U \equiv - \int d^2 x^\perp T_t^t(x^\perp) = -2\pi \int_0^\infty T_t^t r dr \quad \text{and} \quad T \equiv -2\pi \int_0^\infty T_z^z r dr, \quad (11.28)$$

which are, respectively, the energy per unit length and the tension.

In the Abelian Higgs model, the fields cannot depend on the string internal coordinates, so that it is not possible to distinguish between U and T . As a matter of fact, the direct calculation, performed by inserting the solution (11.22) in (11.27) shows that we have

$$\begin{aligned} U &= T = 2\pi \int_0^\infty \left[\left(\frac{d\varphi}{dr} \right)^2 + \frac{1}{2\tilde{q}^2 r^2} \left(\frac{dQ}{dr} \right)^2 + \frac{\varphi^2 Q^2}{r^2} + \lambda (\varphi^2 - \eta^2)^2 \right] r dr, \\ &= 2\pi\eta^2 \int_0^\infty \left[\left(\frac{dX}{d\rho} \right)^2 + \frac{1}{\tilde{q}^2 \rho^2} \left(\frac{dQ}{d\rho} \right)^2 + \frac{X^2 Q^2}{\rho^2} + \frac{1}{8} (X^2 - 1)^2 \right] \rho d\rho. \end{aligned} \quad (11.29)$$

This function is thus a function of \tilde{q} only, multiplied by η^2 , and is represented in Fig. 11.8.

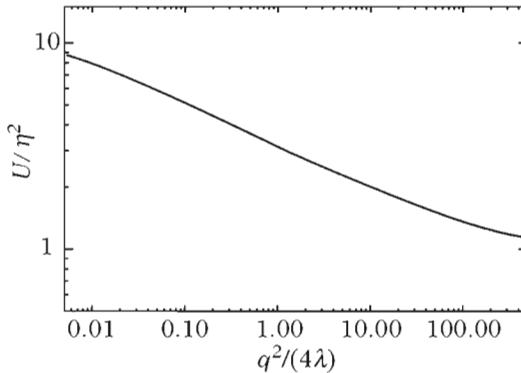


Fig. 11.8 Energy per unit length of a vortex, in units of the Higgs field vacuum expectation value, U/η^2 , as a function of the dimensionless parameter $\tilde{q}^2 = q^2/(4\lambda)$ in the Abelian Higgs model.

11.3.2.2 Nambu–Goto strings and superconducting strings

The fact that the energy per unit length and the tension are equal arises from the fact that the string has no structure. As a result, these strings have a Lorentz symmetry along their direction: we call them Nambu–Goto strings [10].

On some strings, currents can propagate. This depends on the coupling of the fields constituting the string with the other fields contained in the realistic theory. We can then show that the currents propagating are of the superconducting type [11]. Detailed studies [12] of these superconducting strings show that the relation between U and T , which is *the equation of state*, depends non-trivially on a parameter, called the state parameter⁶ [13]. These strings can induce important effects in cosmology and cause catastrophes similar to that of the monopoles since there exist equilibrium configurations, called vortons [14], whose density in the Universe turns out to be only compatible with observations if the energy scale of the symmetry breaking that forms these strings is very low compared to E_{GUT} , in practice by many orders of magnitude, with the exact value depending on the precise model considered [15].

11.3.2.3 Global strings

An interesting special case of the previous model is that for which the coupling constant $q \rightarrow 0$, i.e. where the $U(1)$ symmetry is global. The strings obtained are then called global strings, and in this case, the Goldstone boson cannot be absorbed by the gauge field since the latter is now absent from the theory. There is thus a massless scalar particle that can propagate the interaction, and therefore be responsible for a long-range interaction. This implies first that the total energy of the vortex is no longer confined to a finite region: in fact, it turns out that the energy density of the string

⁶Denoted by w , the state parameter of a superconducting cosmic string is related to the amplitude of the current flowing along the string, while its sign indicates whether this current is space-like, time-like, or even light-like (for $w = 0$).

diverges logarithmically when moving away from the centre: $U_{\text{global}} \propto \ln(r/r_h)$ when $r \gg r_h$, r_h being the Compton wavelength of the Higgs field.

This divergence implies that the total energy of global strings is expressed by $E \sim L\eta^2 \ln(L/r_h)$ for a loop of length L . Moreover, because of the long-range interaction, the energy loss due to the emission of Goldstone bosons (Goldstone radiation) is important. It has been estimated [16] that an entire loop contained in the Hubble radius would lose all of its energy in the form Goldstone radiation in a time of the order of 20 times its length. This process is thus extremely efficient.

Such a Goldstone boson naturally appears in many semi-realistic approximate versions of string theory low-energy compactification, where they can, for instance, play the role of the axion (see Chapter 7). As a consequence, this radiation can give constraints on these high-energy theories via their consequences at very low energy.

11.4 Monopoles

A symmetry-breaking scheme capable of producing monopoles is $\mathcal{G} = \text{SO}(3)$ broken into $\mathcal{H} = \text{U}(1)$. This breaking can arise through a Higgs field

$$\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}$$

in the representation **3** of $\text{SO}(3)$, with $\phi_a \in \mathbb{R}$. The Lagrangian of such a model is then

$$\mathcal{L}_{\text{monopole}} = -\frac{1}{2} |D_\mu \phi_a|^2 - \frac{1}{4} (F_{\mu\nu}^a)^2 - \lambda (\phi_a \phi^a - \eta^2)^2, \quad (11.30)$$

with $D_\mu \phi_a = \partial_\mu \phi_a - q\epsilon_{ab}^c A_\mu^b \phi_c$ and $F_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a - q\epsilon_{abc}^a A_\mu^b A_\nu^c$, where we recall that ϵ_{abc} is the completely antisymmetric Levi-Civita tensor.

A so-called 't Hooft–Polyakov monopole configuration [17], is obtained in spherical coordinates, by considering a static radial field solution of the form

$$\phi^a(x) = \varphi(r) \frac{x^a}{r} \quad \text{and} \quad A_i^a = \frac{Q(r) - 1}{qr^2} \epsilon_{ij}^a x^j, \quad A_0^a = 0, \quad (11.31)$$

with boundary conditions similar to that of the string, namely $\varphi(0) = 0$, $\varphi(\infty) = \eta$, $Q(0) = 1$ and $Q(\infty) = 0$. From the ansatz (11.31), the Euler–Lagrange equations read

$$\left\{ \begin{array}{l} \frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{d\varphi}{dr} \right) = \frac{2Q^2\varphi}{r^2} + 4\lambda (\varphi^2 - \eta^2) \varphi, \\ \frac{d^2Q}{dr^2} = q^2 \varphi^2 Q + \frac{Q}{r^2} (Q^2 - 1). \end{array} \right. \quad (11.32)$$

These equations are non-linear in the gauge field due to the non-commutativity of $\text{SO}(3)$. This makes the system just as analytically insoluble as (11.26), and we should thus again resort to a numerical integration to obtain a result. The solutions are very similar to those obtained in the vortex case of Fig. 11.7.

The energy-momentum tensor now has only one independent integrated component, which is normal since in this case we can see the configuration as a point-like particle. We obtain the energy of the monopole, or in other words its mass, by integrating T^t_t ,

$$M_M = \frac{4\pi\eta}{\sqrt{\lambda}} \int r^2 dr \mathcal{R} \left[\varphi(r), Q(r); \frac{q^2}{\lambda} \right], \quad (11.33)$$

where \mathcal{R} is a complicated functional of the fields (whose exact form can be found in Ref. [18]) depending on the single parameter q^2/λ . We obtain the approximate form $E \sim 4\pi\eta (q^2/\lambda)^{-0.53}/\sqrt{\lambda}$, which shows that the monopole mass is almost independent of the Higgs field self-coupling constant.

In the limit $q^2/\lambda \rightarrow \infty$, called the Bogomol'nyi, Prasad and Sommerfield (BPS) limit [19], we obtain an analytic solution, that is

$$\varphi = \eta \left[\cotanh \left(\frac{r}{r_0} \right) - \frac{r}{r_0} \right] \quad \text{and} \quad Q = \frac{r/r_0}{\operatorname{sh}(r/r_0)}, \quad \text{with} \quad r_0^{-1} = q\eta,$$

with the mass $M_M = 4\pi\eta/q$. It is a surprising fact that this limit happens to represent also a good approximation to the solution for any possible value of the parameter (q^2/λ) .

11.5 Textures

This overview would not be complete without mentioning textures, which are, nevertheless, not strictly speaking topological defects since they are not stable in three dimensions. Textures are non-trivial field configurations, with boundary conditions that make them topologically stable only if space has at least four dimensions: this would be the same thing as constructing a topologically non trivial configuration with the action (11.30), but in a plane. This is why these objects are necessarily dynamical [16]. Every study indicates that they are not compatible with cosmological observations, or that they can only have played a minor role. In other words, they do not lead to any interesting constraints. However, as for the other defects, their existence, if not their stability, depends on a symmetry-breaking scheme of the type (11.12) for which the vacuum manifold \mathcal{M} is non-trivial. Note, by the way, that if there are more than three spatial dimensions, as predicted in phenomenological models inspired by superstring theory, then textures, and other new objects of the same type, can become stable and thus play a new role in cosmology. Textures in these theories will be seen in our Universe as defects of smaller dimensionality, walls, strings or monopoles.

11.6 Defects in general

On the basis of the symmetry-breaking schemes (11.12), it is possible to generalize the conditions under which different types of topological defects form. For this, we make use of homotopy groups, which allow us to understand the topological properties of the vacuum manifold, \mathcal{M} , itself seen as a topological space.

11.6.1 Connectedness

The quotient groups \mathcal{M} that are of interest to us, are, in general, manifolds, i.e. spaces locally isomorphic to \mathbb{R}^n , where n is the dimension of the space considered. This makes it possible to resort to the usual definitions of topology, as known in \mathbb{R}^n . We first define connectedness in the following way: \mathcal{M} is connected by arcs if for any pair of points (x_1, x_2) in \mathcal{M} , there is a continuous map α from $[0, 1]$ into \mathcal{M} satisfying $\alpha(0) = x_1$ and $\alpha(1) = x_2$. Such a map is called a path.

Let us now consider two topological spaces X and Y , and two continuous maps α_1 and α_2 from X into Y . These two maps are said to be homotopic if α_2 can be continuously deformed into α_1 , i.e. if there is a continuous map $F : X \times [0, 1] \rightarrow Y$ such that, $\forall x \in X$, $F(x, 0) = \alpha_1(x)$ and $F(x, 1) = \alpha_2(x)$. The set of homotopic maps divides the space of maps from X to Y into equivalence classes that are invariant under the homeomorphisms⁷ of X or of Y . These equivalence classes are thus topological invariants of the pair of spaces X and Y . Comparing systematically a given space with the sphere S^n of the same dimension (reference topological space), it is possible to classify directly any topological space by these equivalence classes [20].

In particular, if X is a connected space and if there is at least one homeomorphism $f : X \rightarrow Y$, then Y is also connected. Indeed, if Y was not connected, we can then write it in the form $Y = Y_1 \cup Y_2$ with Y_1 and Y_2 two open sets satisfying $Y_1 \cap Y_2 = \emptyset$. Since by hypothesis f is a homeomorphism, then its inverse f^{-1} exists and the open spaces $f^{-1}(Y_1)$ and $f^{-1}(Y_2)$ of X must satisfy $f^{-1}(Y_1) \cup f^{-1}(Y_2) = f^{-1}(Y) = X$ and $f^{-1}(Y_1) \cap f^{-1}(Y_2) = \emptyset$. For this, X should not be connected, in contradiction with the initial hypothesis. This is why topological defects, which exist due to properties similar to the connectedness of the vacuum \mathcal{M} , are stable. There is no homeomorphism of space onto itself, such as, for instance, a physical dynamical phenomenon, susceptible to suppressing them and they can only form through physical discontinuities such as phase transitions.

11.6.2 Fundamental group

The notion of connectedness introduces paths that can form loops that are simply closed paths, going from a point x_0 to itself. As for paths, we can say that two loops α and β based in x_0 are homotopic if it is possible to deform one into the other, i.e. if there is a continuous map $H : [0, 1] \times [0, 1] \rightarrow \mathcal{M}$ such that $H(t, 0) = \alpha(t)$, $H(t, 1) = \beta(t)$ and $H(0, s) = H(1, s) = x_0$. We write $\pi_1(\mathcal{M}, x_0)$ the set of equivalence classes of the loops based at the point x_0 of \mathcal{M} .

Let us now consider a loop product law consisting of following first the first loop, and once returning to x_0 , following the second. This is represented by

$$\gamma(t) = \begin{cases} \alpha(2t) & \text{for } 0 \leq t \leq \frac{1}{2} \\ \beta(2t - 1) & \text{for } \frac{1}{2} \leq t \leq 1, \end{cases} \quad (11.34)$$

⁷A homeomorphism is defined as an invertible continuous map with continuous inverse. Acting a homeomorphism β on X or on Y amounts to decomposing the continuous initial map $\alpha : X \rightarrow Y$ in the form $\alpha : X \rightarrow \beta(X) \rightarrow Y$ or $\alpha : X \rightarrow Y \rightarrow \beta(Y)$. Such a decomposition leaves the classification unchanged.

thereby defining $\gamma \equiv \alpha \star \beta$ as the loop product of α and β . The product of the equivalence classes in π_1 , namely $[\alpha]$ and $[\beta]$, is defined in the same way by $[\gamma] = [\alpha] \circ [\beta]$, where $[\gamma]$ represents the equivalence class of the loops of the form $\gamma = \alpha \star \beta$. Adding the identity loop ‘Id’, simply defined by $\text{Id}(t) = x_0$, $\forall t \in [0, 1]$, to this structure, then the set $\pi_1(\mathcal{M}, x_0)$ with the equivalence class product operation and the identity element form a group: the first homotopy group (also called fundamental group) of \mathcal{M} at x_0 .

The fundamental group is defined on the entire space: if \mathcal{M} is connected by arcs, as is the case for most interesting vacuum manifolds, and $(x_1, x_2) \in \mathcal{M}^2$, then⁸ $\pi_1(\mathcal{M}, x_1) \sim \pi_1(\mathcal{M}, x_2)$, that we denote more simply by $\pi_1(\mathcal{M})$. Note, furthermore, that if X and Y are homeomorphic and path-connected, then $\pi_1(X) \sim \pi_1(Y)$. In other words, the first homotopy group is a topological invariant.

11.6.3 Homotopy group

11.6.3.1 Definitions

It is possible to define homotopy groups of arbitrary rank $p \in \mathbb{N}$, which reduces to the fundamental group in the special case $p = 1$. In the same way a loop is defined as a map of the segment $[0, 1]$ into \mathcal{M} with identification of the extremities, we call p -loop a continuous map of the p -dimensional cube into \mathcal{M} in which the contour is identified with a point of \mathcal{M} . For instance, a 2-loop is homeomorphic to the surface of the 2-dimensional sphere S^2 .

One can show [20] that the set of equivalence classes of the p -loops, with a product operation generalizing that defined in (11.34) and a neutral element is still a group: the homotopy group of rank p of \mathcal{M} , denoted by $\pi_p(\mathcal{M})$.

The classification is completed by the homotopy group of order zero, $\pi_0(\mathcal{M})$, which includes the set of disjoint subsets of \mathcal{M} . For instance, a path-connected space is a space for which $\pi_0 \sim \{\text{Id}\}$, while $\pi_0(Z_2) \sim Z_2$.

11.6.3.2 Properties

An interesting property of the first homotopy group is related to the product groups. We have

$$Z = X \times Y \implies \pi_1(Z) \sim \pi_1(X) \oplus \pi_1(Y), \quad (11.35)$$

i.e. the fundamental group of a product of groups is isomorphic to the direct sum of the fundamental groups of each separated group. This property allows us to quickly and easily identify the homotopy groups of the quotient spaces involved in realistic grand unified theories.

An especially useful example, in particular for cosmic strings, is the case where the topological space considered is the gauge group $U(1)$, i.e. the S^1 circle to which it is topologically equivalent (isomorphic). The different loops it is possible to construct on this circle are characterized by the number of times the loop in question winds around the circle knowing that a loop that does not make it completely around can be continuously deformed to a point. Furthermore, if we provide an orientation to the

⁸By $A \sim B$ we denote the fact that A and B are isomorphic.

circle, so that clockwise is inequivalent to anticlockwise, we can completely characterize the equivalence class of a given loop with a relative integer. We thus have

$$\pi_1[U(1)] \sim \pi_1(S^1) \sim \mathbb{Z}. \quad (11.36)$$

This relation is generalizable to the case of the p -dimensional sphere S^p to which the study of the spaces of interest can often be reduced. We find

$$\pi_p(S^p) \sim \mathbb{Z}, \quad (11.37)$$

with all the homotopy groups of lower rank being trivial, i.e. $\pi_k(S^p) \sim \{\text{Id}\}$ for all $k < p$.

11.6.3.3 Classification of defects

Equipped with these definitions and properties, it is now easy to establish a classification of the different defects that can form during a phase transition of symmetry breaking following the scheme (11.12). Table 11.1 summarizes the properties we now derive.

Table 11.1 Vacuum homotopy group and possible defects.

| Defect | Non-trivial homotopy group | Example |
|-----------|----------------------------|----------------------------------|
| Walls | π_0 | $Z_n \rightarrow \{\text{Id}\}$ |
| Strings | π_1 | $U(1) \rightarrow \{\text{Id}\}$ |
| Monopoles | π_2 | $SO(3) \rightarrow U(1)$ |
| Textures | π_3 | $SO(4) \rightarrow SO(3)$ |

Domain walls form when different domains can appear, as in the case of the Z_2 symmetry of model (11.13). Similarly, if the vacuum resulting from the breaking has many distinct elements of the same energy, then domains will form again. In other words, walls will necessarily exist when breaking a discrete symmetry, with a non-trivial zeroth-order homotopy group $\pi_0(\mathcal{M})$.

The case of vortex lines in the Abelian Higgs model provides an example of breaking with a non-trivial fundamental group of the vacuum manifold. This property is easily generalizable: if $\pi_1(\mathcal{M}) \not\sim \{\text{Id}\}$, cosmic strings will result from the phase transition.

Monopoles, on the other hand, are systematically formed as soon as the vacuum second homotopy group is non-trivial, and textures result from the properties of the third homotopy group. We are here constrained by the number of spatial dimensions. If space had more than three dimensions, not only could there exist topologically stable texture configurations, but also defects of higher dimensions, implying homotopy groups of higher degrees. In a physical space of dimension n , only defects with non-trivial groups π_p with $p < n$ can exist.

11.6.4 Semi-topological defects

Every situation studied so far involved the topological properties of the vacuum, i.e. the topology of the group of transformations that leaves the minimum of the potential

invariant. Since this classification does not involve considerations on the kinetic energy, i.e. gradient terms in the Lagrangian, it does not distinguish between global (rigid) and local (gauge) invariances. Nonetheless, there exists a case where this distinction is important, this is when local and global invariances are broken in such a way that the vacuum is trivial, but such that kinetically stable configurations, called ‘semi-topological’ defects, can be formed; this implies global invariances.

An example where these semi-topological defects appear is given by the Lagrangian (11.6) for a doublet, ϕ , of $SU(2)$. The symmetry is then $SU(2)_{\text{global}} \times U(1)_{\text{local}}$ since there is no gauge boson associated with the $SU(2)$ symmetry. During the symmetry breaking, the Higgs field takes a non-vanishing vacuum expectation value and the gauge boson B_μ becomes massive. However, unlike the model where only the $U(1)$ symmetry was involved, there is still a remaining $U(1)_{\text{global}}$ symmetry after the symmetry breaking. The initial symmetry transformation was in fact

$$\phi \rightarrow \phi' = e^{iq\alpha(x)} \phi e^{i\beta \cdot \sigma}, \quad \text{and} \quad B_\mu \rightarrow B'_\mu = B_\mu - \partial_\mu \alpha(x),$$

where β is a constant vector of $SU(2)$, σ the Pauli matrices, and α an arbitrary function of the coordinates.

To minimize the potential, the sum of the squared modulus of the Higgs doublet should be fixed. For this, we can, for instance, choose to cancel one of the components, and fix the modulus of the other. Choosing

$$\phi_0 = \begin{pmatrix} 0 \\ \eta \end{pmatrix},$$

there is still an invariance under the transformations

$$\phi \rightarrow \phi' = \phi \exp \left[i\xi \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \right],$$

ξ being an arbitrary constant. These transformations form a $U(1)_{\text{global}}$ group and the symmetry-breaking scheme is thus here $SU(2)_{\text{global}} \times U(1)_{\text{local}} \rightarrow U(1)_{\text{global}}$. The vacuum \mathcal{M} has therefore essentially the structure of $SU(2)$, which is topologically equivalent to a 3-sphere that has a trivial fundamental group, since it is simply connected. So there should not be any stable strings in this model.

Actually, the vacuum 3-sphere can be seen as the direct product of the sphere S^2 and the circle S^1 , the latter corresponding to the $U(1)$ gauge invariance. We can then study more precisely how it is possible to evolve dynamically from one field configuration to another. Going from one point to another on the 3-sphere does not require any potential energy, since the 3-sphere is by definition a surface of constant potential energy. Similarly, moving along an S^1 circle requires neither potential nor kinetic energy since the initial $U(1)$ invariance is gauged. However, going from one S^1 circle to another on the S^2 sphere does cost kinetic energy. As a consequence, we can imagine field configurations along which ϕ remains on one and only one circle, thus at the same point of $SU(2)$, while forming a vortex line. This will simply be equivalent to the normal structure of a string, but with one component of the doublet fixed to zero, with no possibility of being modified. Such configurations effectively exist, but it is

not certain that they are formed with the same probability as purely topological configurations. Since these semi-topological configurations are only stable for dynamical reasons, the SU(2) symmetry can be gauged, but with a coupling constant sufficiently different from that of the U(1) so that there are noticeable differences between the gauge terms of the different group factors.

The standard electroweak model has exactly the kind of structure explained above, and we thus expect to find strings formed according to this scheme. For the standard model, these strings are only stable if the mass of the Higgs boson is lower than that of the lightest vector boson, in contradiction with experimental constraints. Moreover, because of the actual value of the gauge coupling constants, it turns out that if strings formed in this way were stable, they could be so only under the condition that their winding number is $n = \pm 1$.

11.7 Walls in cosmology

Even though defects are generic in particle-physics theories, in practice they are predicted for phase transitions that happened around the grand unification breaking, i.e. an energy scale inaccessible in accelerators. As a consequence, it is mostly in cosmology that these objects can have some relevance and play a role, which we shall now elucidate. In turn, it is cosmological observations that will allow us to impose constraints on the defects and thus on the high-energy symmetry breaking schemes that predict them.

The case of domain walls is the simplest to treat in the context of cosmology. Taking into account the correlation length ξ on which they form, one expects to find walls of every size in the Universe. The energy per unit surface computed in (11.20) leads to $U_{\text{wall}} \sim \eta^3 \simeq 4,5 \times 10^{48} \text{ g}\cdot\text{cm}^{-2}$, for a GUT scale of order $\eta \sim 10^{15} \text{ GeV}$. For a structureless wall spreading across the Universe, i.e. on a typical size comparable to the Hubble distance H_0^{-1} , this leads to a total energy density $\rho_{\text{wall}} \sim U_{\text{wall}} H_0 \simeq 10^{20} \text{ g}\cdot\text{cm}^{-3}$. The density parameter of the domain-wall component is thus of the order of

$$\Omega_{\text{wall}} \sim 10^{49} \times \left(\frac{\eta}{10^{15} \text{ GeV}} \right)^3 \sim \left(\frac{\eta}{100 \text{ MeV}} \right)^3, \quad (11.38)$$

which requires that no wall-producing phase transition could have occurred at an energy scale larger than 100 MeV, a scale that is actually well tested in accelerators. This first and rough estimate, while quite unrealistic, gives a good indication of the magnitude of the problem.

11.7.1 Distribution and evolution

The constraint (11.38) relies on the hypothesis that there could still exist today a planar domain wall spreading over a distance of the order of the Hubble radius. Actually, right after the phase transition, the characteristic scale of the inhomogeneities along the walls must be of the order of the correlation length, ξ . This length therefore gives an estimate of the mean curvature radius of the wall. As for their distribution in the Universe, this is a percolation problem that goes beyond the scope of this book [21] whose well-known result is the following: shortly after the transition, the system must be dominated by a large wall with an extremely complicated structure spreading along

the entire Universe. The rest is split between a few closed walls of size comparable to the correlation length, the probability to have a wall of larger size being exponentially suppressed.

Since the walls have an important surface density and tension, the closed wall immediately begins contracting, and eventually decays in a short time so that their contribution, which is already weak initially, rapidly becomes completely negligible. The large walls are subject to two effects: the tension force that tends to accelerate them rapidly to relativistic speeds, and the friction due to the existing matter density ρ . If $\langle v \rangle$ is the mean particle speed of the surrounding fluid when they hit the wall, then the force per unit of surface acting on the wall is $\mathbf{F}/S = \Upsilon \rho \langle v \rangle \dot{R}$, where \dot{R} is the speed of a local curvature perturbations of the wall and Υ a numerical coefficient determined by the interaction cross-sections of the wall with the particles from the environment. For a structureless wall,⁹ the wall tension is simply given by U_{wall}/R . The evolution of R can be evaluated in an overdamped regime in which we consider that the friction term is very important so that the wall is practically in equilibrium. We thus postulate a balance between these forces, so that

$$\Upsilon \rho \langle v \rangle \dot{R}(t) = \frac{U_{\text{wall}}}{R(t)}.$$

This regime is reached during a phase dominated by a perfect fluid (in general, radiation), for which the density is $8\pi G_N \rho = 3H^2$, and the scale factor behaves as $a \propto t^\alpha$. This implies

$$\frac{3\Upsilon}{32\pi G_N t^2} \frac{dR}{dt} = \frac{U_{\text{wall}}}{R},$$

whose solution

$$R = \sqrt{R_0^2 + \frac{64\pi G_N U_{\text{wall}}}{9\Upsilon} t^3} \longrightarrow \frac{8}{3} \sqrt{\frac{\pi G_N U_{\text{wall}}}{\Upsilon}} t^{3/2},$$

allows us to deduce the characteristic speed

$$\dot{R} \sim \sqrt{\frac{t}{t_*}}, \quad \text{with} \quad t_* \equiv \frac{\Upsilon}{16\pi G_N U_{\text{wall}}}.$$

Since after a time t_* this speed is of the order of the speed of light, the overdamped regime approximation is no longer valid.

The function $R(t)$ is interpreted as the typical curvature radius of the set of walls present in the Universe at time t , so that the energy density contained in these walls is simply the surface energy multiplied by the area of this surface, compared to the volume it occupies. In other words, we find $\rho_{\text{wall}} \simeq U_{\text{wall}} R^2 / R^3 = U_{\text{wall}} / R$. Dividing by the critical density, we find

$$\Omega_{\text{wall}}^{\circ, d} = \frac{\Upsilon}{4\alpha^2} \sqrt{\frac{t}{t_*}}, \tag{11.39}$$

⁹Some walls have a structure [22] due to some currents that can propagate along them: it is a mechanism similar to that involved for the superconducting strings described below, but less interesting due to the very strong constraints on the possible wall existence.

(where ‘o.d’ stands for overdamped) which shows that walls dominate the Universe as soon as t approaches t_* if $\Upsilon \sim 1$ (this is strictly the case for radiation for which $\alpha = \frac{1}{2}$).

The value of Υ has been estimated in particular for the case of radiation. In grand unified theories, the photon belongs to a gauge boson multiplet whose other members are very massive. Depending on how the symmetry is broken, the photon, i.e. the massless gauge field remaining after the symmetry breaking, is not necessarily the same component of the given multiplet on each side of the wall, on which the probability of interaction of the photon with the wall strongly depends. This question was resolved in 1974 [23]: Everett showed that if the component of the photon on one side of the wall corresponded to a massive vector on the other side, then the reflection probability is large, and the coefficient Υ is of order unity. In the opposite case, the wall is practically transparent to light and $\Upsilon \ll 1$.

Different regimes have been studied besides the overdamped one just discussed, and always with the same result: after a time t_* , the Universe is dominated by the walls. Independently of the value of Υ , t_* is an extremely short time from the cosmological point of view.

11.7.2 Observational constraints

The domination of the Universe by a domain-wall network, or the simple existence of such a wall still today, leads to numerous observable consequences, which are not observed, and thus transform into very serious constraints on the theories predicting their existence. For instance, a set of walls acts, averaging on the components of the stress-energy tensor, as a perfect fluid with an equation of state $\omega_{\text{walls}} = -\frac{2}{3}$, leading to an accelerated expansion $a \propto t^2$. Even the recent observations of the currently ongoing acceleration of the Universe is incompatible with such a value, as discussed in Chapter 12.

The estimate (11.38) of the wall density is actually too crude in terms of global density since it is based on the existence of walls spreading through the entire Universe. Such a configuration is obviously in complete disagreement with the cosmological principle since it would produce a preferred plane. In order for this principle to remain respected Ω_{wall} should be small. But then, since photons have the possibility of interacting with the wall, even if only gravitationally, this would induce temperature anisotropies in the microwave background. Equation (11.38) should thus be replaced by

$$\left(\frac{\Delta T}{T} \right)_{\text{walls}} \sim 10^{-6} \times \left(\frac{\eta}{1 \text{ MeV}} \right)^3, \quad (11.40)$$

indicating that the symmetry-breaking energy scale should not exceed the order of the MeV.

To conclude, apart from special situations, we can say that it is usually accepted that a theory with broken discrete symmetries, and thus producing walls during the corresponding phase transition, must also include some mechanism to destroy these walls in order not to be excluded by cosmological observations. The simplest method to construct a grand unified theory that is satisfying from this point of view, is to make

it so that no walls are produced. This constrains the number of possible theories, although not significantly.

11.8 Cosmological monopoles

Monopoles, like walls, are extremely constrained, and in most cases lead to unavoidable cosmological catastrophes. The difference with walls, is that it does not seem possible to combine grand unification with the observed absence of monopoles.

11.8.1 GUT monopoles are unavoidable

Let us assume that the unification group is \mathcal{G} . We usually choose it such that its first two homotopy groups are trivial: $\pi_1(\mathcal{G}) \sim \pi_2(\mathcal{G}) \sim \{\text{Id}\}$. However, the standard model of particle physics indicates that at low energy, there must remain a local U(1) invariance corresponding to the electromagnetic sector, since experimental constraints show that the photon is massless. Moreover, we can show the implication

$$\pi_n(\mathcal{G}) \sim \pi_{n-1}(\mathcal{G}) \sim \{\text{Id}\} \implies \pi_n(\mathcal{M}) \sim \pi_{n-1}(\mathcal{H}), \quad (11.41)$$

where we have used the conventions (11.12).

Since in the general symmetry-breaking scheme, whatever it is, that leads from the grand unification symmetries to those currently observed, we recover an unbroken U(1) at the end, this means that there is necessarily such a U(1) factor in the breaking sequence. Since U(1) is topologically equivalent to a circle, its first homotopy group is \mathbb{Z} , so that, due to (11.35) applied to π_2 , the second homotopy group must contain a \mathbb{Z} factor. As a result, independently of the grand unified theory, one has

$$\pi_2(\mathcal{M} = \mathcal{G}/\mathcal{H}) \supseteq \mathbb{Z} \not\sim \{\text{Id}\}, \quad (11.42)$$

and the appearance of monopoles is unavoidable.

11.8.2 The monopole problem

The monopole excess problem is not, like in the wall case, due to the mass or to the perturbations of a single monopole, but to their density and their evolution. The only way to make the monopoles produced during the phase transition disappear is to annihilate all the pairs of monopoles and antimonopoles. All the pairs that are initially very close to one another will annihilate relatively rapidly, but the small remaining density must scatter for a long time in the surrounding medium before managing to decay: this is a very slow process.

11.8.2.1 The causality argument

In many references, one finds a causality argument allowing for an estimate of a lower bound on the monopole density present in the Universe. This argument has mainly a historical value since it relies on the existence of a cosmological horizon, and shows the existence of a problem without even effectively computing the real monopole density. Unfortunately, if the horizon problem is resolved by whatever means, then the argument no longer holds, and one should resort to a more complete calculation. We, however, present it for the sake of completeness.

Let us thus assume that the Universe starts its evolution with a singularity at $t = 0$, followed by a radiation-dominated phase, $a \propto \sqrt{t}$. In this case, causality provides an upper bound on the correlation length ξ of the Higgs field responsible for the appearance of the monopoles by $\xi < d_{\text{H}}$, where the horizon size d_{H} is given by (3.107).

Using (4.19), we thus obtain

$$\xi < \sqrt{\frac{45}{4\pi^3 g_* G_N}} T_c^{-2},$$

where T_c is the temperature at the phase transition or that when the bubble nucleation starts, depending on the order of the transition. Let p be the probability of a non-trivial field configuration in a correlation volume; p mainly depends on the topological structure of the vacuum and is, for most theories studied so far, a number of order 10^{-1} . Since the initial density of monopoles produced at the transition is of order $n_{\text{M}} \sim p\xi^{-3}$, we find an initial density

$$\left. \frac{n_{\text{M}}}{T^3} \right|_{\text{ini}} = \frac{n_{\text{M}}(T_c)}{T_c^3} \gtrsim p \left(\frac{4\pi^3 g_* G_N}{45} \right)^{3/2} T_c^3 \sim 10^{-10},$$

which is too high to be sufficiently reduced by the evolution, as will be seen below.

11.8.2.2 Initial density

The real initial density can be computed in a more efficient way without resorting to the causality argument, especially since this argument would have had no role to play if the horizon problem was resolved before the phase transition.

The density of defects produced depends on the order of the transition. In particular, we can find values of the parameters for which the initial density is as small as we want if the transition is strongly first order. This is due to the fact that, for these parameters, the temperature should be very small compared to the critical temperature before the nucleation starts to be efficient. In this case, the energy contained in each bubble can be quite small, whereas the mass of the monopole that forms remains the same, and we thus expect that the probability of formation is reduced. Such a first-order phase transition gives a first solution to the monopole problem.

If the transition is second order, then close to the transition, the fluctuations of the field responsible for the appearance of the monopoles are very important and many defects appear, quickly eliminated by pair annihilations until the temperature drops to the level where the fluctuations are negligible. This gives a distribution of monopoles that will rapidly evolve by annihilation of neighbouring pairs. Indeed, when a monopole and an antimonopole are put in contact, before annihilating, they form a bound state called monopolonium [24] similar to a hydrogen atom but with the electrostatic interaction replaced by a magnetostatic interaction. Unlike the hydrogen atom, this monopolonium is formed by two particles with structure and are capable of annihilating each other since the topological stability condition of the monopole is canceled by that of the antimonopole. The closer the initial pair, the faster does the monopolonium decay and in the initial conditions, we can neglect this characteristic time.

After the first annihilations, there remains a diluted distribution with correlation length larger than the radiative capture length (considered as being the dominant mechanism to form pairs susceptible to decay). Actually there are other, more efficient, mechanisms [25], but the conclusions remain unchanged if we take them into account and they finally only complicate the calculations. At a temperature T , the capture distance of a monopole by an antimonopole is given by $d_c = g^2/(4\pi T) \simeq 75/T$, where g is the monopole magnetic charge, of the order of $2\pi/q$, and q the reduced fundamental electric charge in the theory considered [i.e. $q^2/(2\pi) \simeq \frac{1}{137}$]. The distribution we are thus interested in has, right after the phase transition that produced it, a monopole number density per unit volume $n_M^0 \sim (4\pi/g^2)^3 T^3 \sim 10^{-6} T^3$, equal to that of the antimonopoles.

11.8.2.3 Evolution of a gas of monopoles

The evolution of this distribution, which is assumed to be isotropic, is subject to two effects: the annihilation between monopoles and antimonopoles and the dilution under the effects of the expansion, and both these effects cancel. The equation governing the evolution is thus [26, 27]

$$\frac{dn_M}{dt} + 3Hn_M = -Dn_M^2, \quad (11.43)$$

where the diffusion coefficient D characterizes the annihilation (this is similar to our study of Chapter 4).

At high temperatures, the mean free path, λ_M , of the monopoles is greater than d_c , and the annihilation rate is simply given by the monopole flux F in the primordial plasma, which is obtained by the relation $F = g^2 n_M \tau / M_M$, where τ represents the characteristic time between two collisions during which the monopole is scattered by an important angle and M_M is the monopole mass. The diffusion coefficient D turns out to be the flux of monopoles divided by the distribution, $D = F/n_M$.

During a collision between a light particle of the plasma, having an energy essentially equal to the plasma temperature T and charge q , and a monopole of magnetic charge g , the scattering cross-section is

$$\sigma \simeq \left(\frac{gq}{4\pi T} \right)^2.$$

The monopole is very weakly scattered at each collision, so that there should be around M_M/T collisions before the monopole scattering is of consequence. From the relation of Table 4.4, the number density of particles of each species on which the monopoles can scatter is $n_{\text{plasma}} = 3\zeta(3)T^3/4\pi^2$, so that the characteristic time after which a monopole is deviated is $\tau = (M_M/T)(n_{\text{plasma}}\sigma)^{-1} = M_M/(BT^2)$, where $B = \frac{3}{4\pi^2}\zeta(3)\frac{g^2}{4\pi} \sum_i q_i^2 \sim 10$ (the sum over the squared charges taking into account the actual number of degrees of freedom g_* at that time), and we thus find

$$D = \frac{g^2}{BT^2}. \quad (11.44)$$

This relation is valid until $\lambda_M \sim d_c$.

Once the time τ is known, the mean free path can be calculated as $\lambda_M = \langle v_M \rangle \tau$, where $\langle v_M \rangle \sim \sqrt{T/M_M}$ is the thermal speed of the monopoles. As a result, the temperature T_f below which the coefficient D is no longer well estimated by (11.44) turns out to be given by $T_f = M_M B^{-2} (4\pi/g^2)^2 \sim 10^{-6} M_M$.

The dominant mechanism is then that of thermal capture. Without going into too much detail, we find [25] that the cross-section for this process is $\sigma \propto T^{-7/5}$, and then $D = \sigma \langle v_M \rangle \propto T^{-9/10}$. If other mechanisms play a role, one can find a different power law, but still with exponent smaller than unity. In order to study the general case, we will thus set

$$D = \frac{A}{M_M^2} \left(\frac{M_M}{T} \right)^p,$$

where A is a (dimensionless) numerical factor taking into account the microphysics of the collision between the monopoles. The evolution of their distribution strongly depends on the value of p .

We now assume that the evolution above occurs during a phase for which the temperature T decreases with the scale factor as $T \propto a^{-1}$, so that $\dot{T}/T = -H$. The Hubble factor is expressed by (4.19), and using (4.21) between the time and the temperature, we can rewrite (11.43) in the form

$$\frac{d}{dT} \left(\frac{n_M}{T^3} \right) = \frac{AC}{M_M^2} \left(\frac{M_M}{T} \right)^p \left(\frac{n_M}{T^3} \right)^2, \quad (11.45)$$

with $C^2 = 4\pi^3 g_* G_N / 45$, and where we note that the term linear in n_M disappears. Relation (11.45) can be integrated and we obtain

$$\frac{T^3}{n_M} = \frac{T_{\text{ini}}^3}{n_M^0} + \frac{AC}{(p-1)M_M} \left[\left(\frac{M_M}{T} \right)^{p-1} - \left(\frac{M_M}{T_{\text{ini}}} \right)^{p-1} \right], \quad (11.46)$$

where T_{ini} is the temperature at which the monopole distribution starts evolving as (11.43); T_{ini} is of the order of T_c , the temperature of the phase transition.

Let us consider first the case of (11.44) with $p = 2$. The solution (11.46) indicates that the annihilations never stop, and the evolution leads to

$$\frac{n_M}{T^3} \sim \frac{B}{g^2 C} T \rightarrow \frac{1}{Bg^2} \left(\frac{g^2}{4\pi} \right)^2 \frac{M_M}{C} \sim 10^{-10}, \quad (11.47)$$

independently of the initial conditions as long as $T \ll T_{\text{ini}}$, which is indeed the case if $T_{\text{ini}} \sim M_M/g$ and $T \sim T_f$, corresponding to the second part of this equation.

From there, the ratio n_M/T^3 never evolves since $p < 1$. However, this phase includes that during which nucleosynthesis must occur, which we know cannot be dominated by anything other than radiation. Actually, in order that monopoles do not dominate the Universe at the nucleosynthesis, we should have

$$\frac{n_M}{T^3} \Big|_{T=1\text{MeV}} \lesssim 10^{-19} \left(\frac{M_M}{10^{16}\text{GeV}} \right)^{-1},$$

in contradiction with the previous prediction.

11.8.3 Possible solutions to the monopole problem

Many solutions to the monopole cosmological excess have been proposed. The most straightforward is to assume that the grand unification phase transition did not produce any monopole. While not impossible, this is a hard-to-believe solution unless the transition was strongly first order with parameters tuned so that the initial production rate is indeed extremely weak. It can also be that the transition never happened and thus that the symmetry was always broken [28]. In this case, the parameters of the theory should again be tuned in order to ensure that the thermal monopole production does not induce the same problem [29]. All of these hypotheses require very precise values of the parameters: this is a possible solution but, in general, not considered as a very satisfying one.

11.8.3.1 Multiple defects

A case not yet discussed so far is that of hybrid defects, i.e. composed either of a wall bounded by a string, or a string ending with two monopoles (or rather a monopole and an antimonopole), that can appear if two homotopy groups of the vacuum manifold are non-trivial. A theory can also have walls and monopoles, hence producing multiple defects, but not hybrid. Finally, there are cases where all three first homotopy groups are non-trivial, and all possible defects can be produced.

When walls and monopoles are produced, we can imagine that the walls, during their motion in the Universe, interact with the monopoles by ‘sweeping’ them away [30]. When a monopole enters a wall, it becomes topologically unstable, in the sense that the Higgs field inside the wall where the symmetry is restored must vanish. The monopole thus decays inside the wall. In order not to replace the monopole problem by the more severe wall problem once all the monopoles initially present have been absorbed, the walls themselves must then also disappear. This is possible as long as the discrete symmetry that produces them is only approximate. The simplest model one can think of, based on the unification group $SU(5)$, already allows for a coherent generalization of this scenario. We can also imagine that strings form later on, creating holes in the walls. These holes then grow until completely dissipating the walls energy into radiation. In this last case, not only are the monopole and wall problems resolved, but we also obtain a string network that can contribute to the primordial perturbations.

Another idea, proposed by Langacker and Pi [31], is to imagine a breaking scheme in which the $U(1)_{\text{elec}}$ symmetry spontaneously breaks at a given time. During this phase, monopoles and antimonopoles are necessarily related by local strings formed from the Higgs that breaks the $U(1)_{\text{elec}}$ and the photon. This has the effect of greatly increasing the pair decay probability since it is now the string tension that attracts the two extremities towards each other. These mechanisms are discussed in detail in Ref. [16].

11.8.3.2 Inflation

All these solutions have a major inconvenience, at least from the perspective of most cosmologists. They depend on the explicit symmetry-breaking scheme, and thus on the representations of the Higgs fields present in Nature, as well as on the values of the microscopic parameters. In that respect, the solution proposed by inflation, having the

advantage to solve the problem in a way independent of the underlying microscopic physics, seems especially privileged. Besides, it is economical, solving also at once the other cosmological puzzles of the standard hot Big-Bang scenario (Chapter 8).

In the inflation case, the entire observable Universe originates from a volume containing at most a few monopoles. If the reheating temperature is sufficiently low, which is the case in models accepted as being compatible with observational data, then the thermal production remains bounded to an acceptable rate.

We can also consider the case that monopoles are produced after the period of inflation, if some conditions are respected, such as the fact that the transition is strongly first order. There are then strong constraints on the unknown parameters, explaining why it is considered as more satisfying to simply resort to a phase of inflation.

11.9 Cosmic strings

The only topological defect not producing a cosmological catastrophe is the vortex line. Actually, unlike the other defects, strings can reach a ‘scaling’ regime (explained later) in the Universe, so that they never risk dominating the global matter content provided their initial contribution did not dominate already. Furthermore, they can produce metric fluctuations susceptible of serving as sources for the formation of the large scale structure. For a long time, these models were considered as potential competitors to inflation. Nowadays, it has been shown that, alone, these models could not reproduce observations, and in particular that of the cosmic microwave background. Cosmological strings therefore appear more as a logical possibility testable in high-precision measurements (a non-inflationary component in the cosmic microwave background, for instance).

11.9.1 General properties

11.9.1.1 Realistic schemes producing strings

Many realistic symmetry-breaking schemes in grand unified theories have been studied, taking into account most of the constraints known from both cosmology and particle physics. In particular, we assume that inflation occurs after the symmetry breaking producing monopoles, and ends with a new phase transition: this is hybrid inflation (see Chapter 8).

Of the set of all possible schemes, only 34 satisfy the cosmological constraints if the initial symmetry is $SO(10)$, and 1124 for E_6 . All these schemes, with no exception, lead to the presence of topological stable strings! This indicates that cosmic strings, despite their weak contribution to the cosmic microwave background, are a generic prediction of grand unified theories. From this, we can hope to constrain the later observationally [32].

As shown in the previous chapter, supersymmetric grand unified theories often need a Z_2 subgroup, the R -parity, that remains unbroken. This symmetry is necessary so that the proton has a sufficiently long lifetime. The general scheme we expect is thus

$$\mathcal{G} \rightarrow \dots \rightarrow SU_c(3) \times SU_L(2) \times U_Y(1) \times Z_2. \quad (11.48)$$

The property (11.35) applied to the homotopy group of order 0 implies that the resulting group contains a factor $\pi_0(Z_2) \sim Z_2 \not\sim \{\text{Id}\}$. From (11.41), we have $\pi_1(\mathcal{M}) \supseteq Z_2 \not\sim \{\text{Id}\}$, and therefore find that strings are unavoidable in this framework.

Finally, strings of cosmological dimensions have been suggested as a plausible prediction of superstring theory (see Chapter 13).

11.9.1.2 Thin strings

A cosmic string is an object extended in one direction, justifying its name, whose width δr is of the order of magnitude of the Compton wavelength associated to the mass of the Higgs field, $\delta r \simeq m_h^{-1}$. This radius, in the case of a string arising from a grand unification symmetry breaking, is of the order of 10^{-28}cm , to be compared to the cosmological dimensions on which strings evolve. It is therefore appropriate to neglect the string width altogether. We thus approximate the string as a Dirac distribution $\delta[x^\mu(\xi^a)]$, where the points $x^\mu(\xi^a)$ describe the string trajectory, the worldsheet, as a function of two intrinsic coordinates, one time-like, $\xi^0 \equiv \tau$, equivalent to the proper time necessary for the description of a free particle trajectory, and the other space-like, $\xi^1 = \sigma$, to indicate the spatial location on the string.

In the thin-string approximation, the total 4-dimensional energy-momentum tensor $T^{\mu\nu}$ of the string can then in all generality be written in the form

$$T^{\mu\nu}(x^\alpha) = \int d^2\xi \tilde{T}^{\mu\nu} \delta^{(4)}[x^\alpha - x^\alpha(\xi^a)]. \quad (11.49)$$

In the neighbourhood of a string, it is always possible to consider it as locally straight and aligned along an axis, z say, in the cylindrical coordinates (r, θ, z) of section 11.3.2. Such a string is described by the energy-momentum tensor (11.27), which becomes, in the small-width limit,

$$T^{tt} = U\delta(x)\delta(y) \quad \text{and} \quad T^{zz} = -T\delta(x)\delta(y),$$

all other components vanishing.

11.9.2 Gravitational effects of strings

The stress-energy tensor for a cosmic string derived above eventually serves as a source for the Einstein equations. This permits calculation of the gravitational effects of cosmic strings.

11.9.2.1 Metric and deficit angle

Far away from the string, we can consider the vacuum solution with axial symmetry, described by the metric

$$ds^2 = - \left(\frac{r}{r_h} \right)^{2\alpha} dt^2 + dr^2 \left(\frac{r}{r_h} \right)^{2\beta} dz^2 + \gamma r_h^2 \left(\frac{r}{r_h} \right)^{2\delta} d\theta^2, \quad (11.50)$$

where r_h is the effective width of the string. The exponents α , β and δ satisfy the constraints

$$1 + \alpha + \beta + \delta = \alpha^2 + \beta^2 + \delta^2 = 1,$$

where at this level, γ is an arbitrary constant.

Using the relation $\nabla^2 \ln(r/r_h) = 2\pi\delta(x)\delta(y)$ in two dimensions, it is possible to solve the linearized Einstein equations (1.126) for the string energy-momentum tensor. Assuming that the typical energies are of the order of the grand unification scale E_{GUT} then $G_N U \sim G_N E_{\text{GUT}}^2 \sim (E_{\text{GUT}}/M_P)^2 \lesssim 10^{-6}$. Identifying this linear solution with the general solution (11.50), we find, to first order in $G_N U$, [33]

$$\alpha = -\beta = 2G_N(U - T) + \dots, \quad \delta = 1 + \dots \quad \gamma = 1 - 4G_N(U + T) + \dots,$$

where the ellipses represent terms of order $(G_N U)^2$.

For a Nambu–Goto string, $U = T$, we find a metric that is mostly that of Minkowski up to the fact that $\gamma = 1 - 8\pi G_N U \neq 1$. This is interpreted by renormalizing the angular variable, i.e. by setting

$$\bar{\theta} = \sqrt{1 - 8G_N U}\theta \simeq (1 - 4G_N U)\theta,$$

transforming $\gamma d\theta^2$ into $d\bar{\theta}^2$, as for a spatially flat metric. Since θ varies between 0 and 2π , $\bar{\theta}$ varies in the interval $0 \leq \bar{\theta} \leq 2\pi(1 - 4G_N U)$. The metric is conical: it is flat, but because of the range of variations of $\bar{\theta}$, it has a missing angle $\delta\bar{\theta} = 8\pi G_N U$.

Considering two circles surrounding the strings and located at the radii r and $r + dr$ then the difference of the circumferences between the two circles is not $2\pi dr$ as for an ordinary flat space, but

$$\varpi(r) = 2\pi \left(1 - \frac{1}{\sqrt{g_{rr}}} \frac{d\sqrt{g_{\theta\theta}}}{dr} \right) = 2\pi \left[1 - \sqrt{\gamma} \delta \left(\frac{r}{r_h} \right)^{\delta-1} \right] \simeq 4\pi G_N(U + T),$$

independent of the distance r at this order of approximation. The angle ϖ is called the deficit angle.

An important point of definition: in general, we consider strings with no inner structure, as is the case of the Nambu–Goto string, or we neglect this internal structure. The energy per unit length and the tension are two numbers with the same order of magnitude, and the deficit angle is

$$\mu \equiv \frac{1}{2}(U + T) \implies \varpi \sim 8\pi G_N \mu.$$

(11.51)

This is the reference quantity in what follows and in most works published on strings.

11.9.2.2 Gravitational lensing

The presence of a deficit angle around a string leads to gravitational mirages that manifest themselves in the form of double images (Fig. 11.9). We can explicitly compute the light-ray trajectories (the geodesics) passing close to a string and deduce the deflection angle Δ (see Chapter 7). We find [33]

$$\Delta = 4\pi G_N \mu + \mathcal{O} \left[\left(G_N^2 \mu^2 \ln \frac{b}{r_h} \right) \right],$$

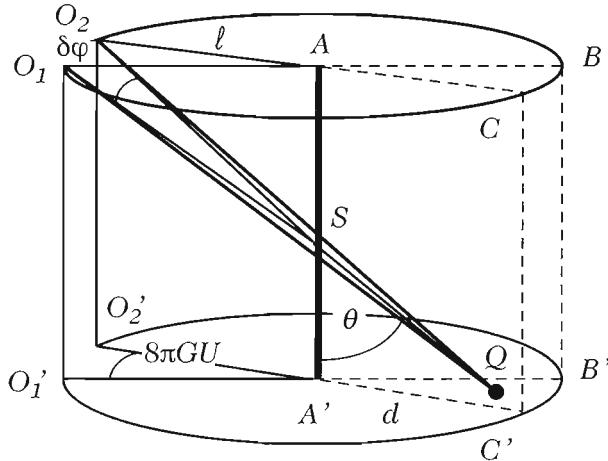


Fig. 11.9 Missing angle around a straight string. The string is located along the axis AA' and we identify the planes $AA'O_2O'_2$ and $AA'O_1O'_1$, separated by an angle $8\pi G_N U \sim 8\pi G_N \mu$. An observer located at the point Q will thus see two rays of light arising from O_1 and O_2 , which, however, represent a single source since the planes are identified. A single source is thus seen as a double image.

where b is the impact parameter; we see that to first order in the approximation, the deflection does not depend on b . The deficit and deflection angles are identical up to a factor 2 in the case of a Nambu–Goto string since

$$2\Delta - \varpi = 4\pi G_N (U - T).$$

In practice, the deficit angle is very small. For an energy of the order of the grand unification scale, we thus find $\Delta \sim 10^{-5}$, that is around $2''$. However, it is not excluded to detect this effect and, as a matter of fact, observations of some galaxy alignments have repeatedly been interpreted in these terms [34]; up to now, these interpretations have failed because of more precise (spectral in particular) data. Referring to Fig. 11.9, we see that the angular distance $\delta\varphi$ between both images, i.e. the angle formed by the sector $O_1\widehat{Q}O_2$, is given by

$$\delta\varphi = 8\pi G_N \mu \frac{\ell}{d + \ell} \sin \theta,$$

where $\ell = AO_1$ is the distance between the string and the source, $d = QA'$ that between the string and the observer, and θ the angle between the line of sight and the tangent to the string.

To go beyond the simple string model, fluctuations in the energy density and geometry along the string have been modelled by assuming that the linear energy density was described by a Poisson random density [35]. New gravitational phenomena are involved, such as the appearance of critical lines where the amplification is infinite (Chapter 7). We can show that the caustic line length per unit of string length is $4\sqrt{3}E(\frac{3}{4})$. These effects are represented in Fig. 11.10.

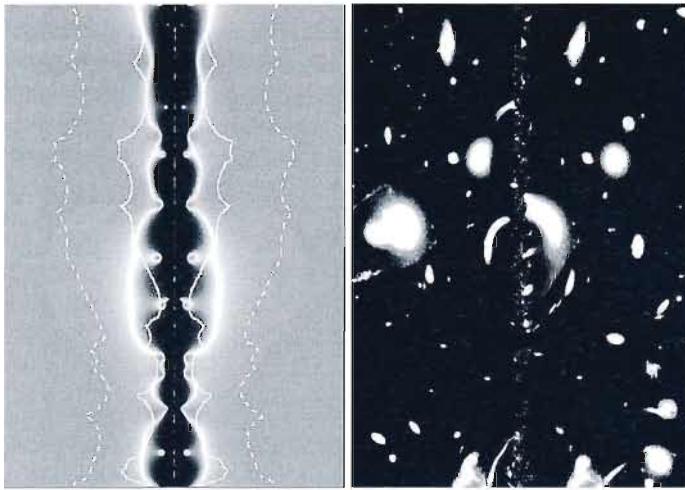


Fig. 11.10 Simulation of gravitational lensing effects by a cosmic string. Effects from string density and geometry fluctuations are taken into account. (left): Amplification map of a Poisson string. The brightest pixels correspond to a greater amplification. The white zone corresponds to a critical line where the amplification is infinite. The continuous white line is the caustic line and the dotted line corresponds to the counterimage of the critical line (Chapter 7). (right): This distortion field is applied to the cluster A2218 arbitrarily localized at $z = 1$. The string is located at $z = 0.8$. The resolution is of 0.1 arcsec. The pair separation is of about 5 arcsec. We note the appearance of a multitude of small images along the string and great arcs along the critical lines. These effects are absent for a smooth string. From Ref. [35].

11.9.2.3 Kaiser–Stebbins effect

Gravitational deflection provides another way to potentially detect a string. We assume that such a string passes between an observer and a light source, and denote by v the string velocity with respect to the observer. Due to the deficit angle, once the string has passed, the source sees its speed modified with respect to the observer so that, by Doppler effect, the frequency ν of the emitted radiation is changed by a factor $\delta\nu$. One finds $\delta\nu/\nu \propto 8\pi G_N \mu \gamma$, where $\gamma = \sqrt{1 - v^2}$ is the string Lorentz factor.

Applying this argument to a thermal distribution of photons such as the cosmic microwave background blackbody, we then see that each frequency is changed in the same relative way. The total spectrum is thus simply shifted by the passage of the string. As a consequence, the temperature changes and we find [36]

$$\boxed{\frac{\Delta T}{T} = 8\pi G_N \mu \mathbf{n} \cdot (\mathbf{v} \wedge \mathbf{t}),} \quad (11.52)$$

where \mathbf{n} is a unit vector along the line of sight and \mathbf{t} a unit vector tangent to the string.

11.9.3 Effect of a network on the microwave background

The Kibble model is used to simulate string networks numerically: one constructs a network whose meshes are separated by a correlation length, and on each mesh, a phase is randomly generated. Studying the repartition of these phases at the network points permits us to determine if there is a cosmic string passing through these meshes. Thanks to this method, it has been shown that roughly 80% of the strings that form during a phase transition are infinite strings, i.e. spreading from one part to another of the volume of the simulation, and that the loop distribution is scale invariant, i.e. the number of loops formed in a spherical shell between the radius r and thickness dr is proportional to dr/r^4 and independent of the correlation length.

11.9.3.1 Initial conditions

To perform numerical simulations of a cosmic-string network, we consider initial conditions imposed long after the phase transition happened, when the temperature has dropped enough so as to be able to neglect the Higgs field thermal fluctuations, for second-order transitions. The distribution of the phases is thus considered as frozen, and one can evaluate the defect distribution simply by evaluating the random distribution of the field on distances larger than the transition correlation length ξ . To obtain this distribution, one uses the algorithm first proposed by Vachaspati and Vilenkin [37]: at each point on a cubic grid with step ξ , a phase is randomly generated between three discrete values, 0, $2\pi/3$ and $4\pi/3$. A cosmic string crosses a square in four points (a plaquette) if the field phase winds by 2π around the plaquette. One thus obtains elementary cubes with entering and outgoing strings. Imposing the flux conservation then permits knowing how to connect them.

Figure 11.11 shows an initial configuration generated by this algorithm, similar to those used in the simulations of Ref. [38]. There are two physical parameters appearing in the simulation: the initial string density, given by the correlation length and the initial volume (as many Hubble volumes as possible at the initial simulation time), and the mean velocity of the string segments of length ξ . Since the initial volume is expanding, the simulation is necessarily limited in its dynamics by the finite size of the numerical grid. Indeed, when the Hubble radius becomes comparable to the size of the initial grid, boundary effects start dominating and the simulation ceases to be physically meaningful. In this grid, on which periodic boundary conditions are imposed (which actually makes the grid a 3-torus), strings that do not close in the volume are said to be infinite or long, and the others are called loops.

11.9.3.2 Reconnection and intercommutation

When two strings meet, there are two topological possibilities: either they cross each other, or they exchange the extremities that meet. Only the dynamics of the microscopic fields composing the strings allows us to determine the appropriate interaction, and simulations have been made indicating that it is the second option that occurs in almost every case. We then say that the strings reconnect, or intercommute.

The most important consequence for cosmology is the constant formation of loops that can dissipate the energy of the network, initially mainly contained in the long strings. As indicated in Fig. 11.12, when two strings meet in two points or when a

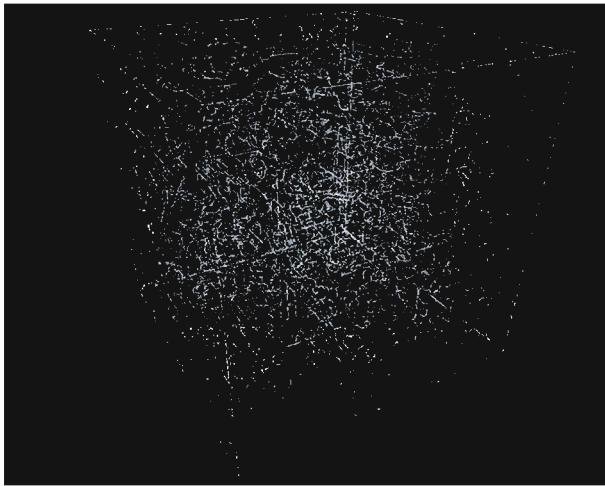


Fig. 11.11 Initial configuration of a string network, computed with the Vachaspati and Vilenkin algorithm [37], in an initial comoving volume of side 18ξ with 26 points per correlation length. Figure produced and provided by C. Ringeval thanks to the original numerical simulation code by D. Bennett and F. Bouchet.

string, due to its motion, happens to fold over itself, the intercommutation mechanism always produces a loop. At each exchange point, the final string shows a very sharp edge. This configuration, called a *kink*, propagates at the speed of light along the string, emitting a large quantity of gravitational radiation.

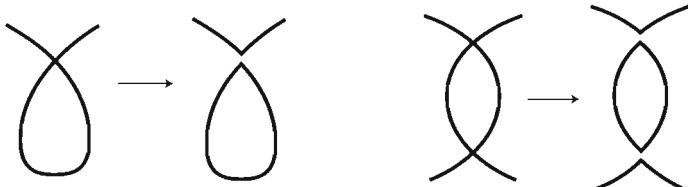


Fig. 11.12 Formation of cosmic-string loops by the intercommutation mechanism and appearance of *kinks*.

11.9.3.3 Scaling

Many analytical models have been proposed to describe the evolution of a string network, first with one length scale [39], then two or three [40] as well as a length scale and a speed [41]. All these models, as well as their numerical simulations, indicate that the string network quickly reaches a special regime, called the *scaling* regime, in which its statistical properties become time independent. Therefore, independently of the precise time at which the phase transition leading to the strings occurred, the

cosmological consequences of these models are expected to be mostly independent of the details of the underlying theory.

The one-scale model allows us to understand how the scaling regime is reached. The scale in question is again the correlation length ξ , that represents as much the characteristic scale between each string as the strings mean curvature (more elaborated models give different values to these lengths). Let us first consider a network of infinite strings, with characteristic length scale denoted by $L_\infty(t) \propto a(t)$. At the time of transition, namely for $t = t_{\text{ini}}$, one has $L_\infty(t_{\text{ini}}) = \xi$, so that in a volume V , the number of strings is $N_{\text{strings}} = VL_\infty^{-3}$. Since the energy of the strings is $E_{\text{strings}} = N_{\text{strings}}L_\infty U$, we find an energy density $\rho_\infty = UL_\infty^{-2} \propto a^{-2}$, implying that a network of infinite strings with no interactions always ends up by dominating over the matter and radiation components. We also point out that this evolution gives rise to an effective equation of state $w_{\text{strings}} = -\frac{1}{3}$ for these strings: such a set of strings is exactly at the border between an accelerating and decelerating solution, and mimicks a curvature term.

Fortunately, this is not the end of the story since strings interact via reconnection and loop formation. Hence, a fraction of the energy is transferred from the system of infinite strings to loops, that eventually dissipate the long string energy, mainly in the form of gravitational radiation, thus preventing domination of the string network. This fraction is proportional to the number of interactions. Assuming that the strings mean speed is of the order of the speed of light (simulations give $\langle v_{\text{strings}} \rangle \sim 0.7$), then the mean time between two interactions is $\sim L_\infty$, and therefore the number of interactions per unit of time and volume is $\propto L_\infty^{-4}$. The fraction of energy transferred from the system of infinite strings into loops of length L_∞ during a time δt is thus $\delta\rho_{\infty \rightarrow \text{loops}} \simeq L_\infty^{-4}\delta t \times UL_\infty$. Taking the expansion of the Universe into account, we find an evolution dictated by

$$\frac{d\rho_\infty}{dt} + 2H\rho_\infty = -\frac{\rho_\infty}{L_\infty}, \quad (11.53)$$

and as a consequence

$$\frac{dL_\infty}{dt} = HL_\infty + \frac{1}{2},$$

whose general solution is

$$L_\infty(t) = a(t) \left[C + \frac{1}{2} \int^t \frac{d\tau}{a(\tau)} \right].$$

Considering a phase dominated by a perfect fluid with equation of state w , and thus a scale factor $a(t) = a_0(t/t_0)^{2/[3(1+w)]}$, we find

$$L_\infty(t) = Ca_0 \left(\frac{t}{t_0} \right)^{2/[3(1+w)]} + \frac{3}{2} \frac{1+w}{1+3w} t,$$

which quickly leads to $L_\infty \propto t$ in the case of a non-accelerating Universe ($w \geq -\frac{1}{3}$). The corresponding energy density then scales as $\rho_\infty \propto t^{-2} \propto a^{-3(1+w)}$, i.e. the same

behavior as the dominant fluid.¹⁰ In other words, the energy density of the string network compared to the density of the Universe rapidly tends towards a constant. This is what we call a scaling regime.

11.9.3.4 Numerical simulations of string networks

From the initial configurations discussed above, it is possible to numerically solve the equations of motion of the strings making the network, taking into account intercommutations. For this, one should first choose the value of the equation of state of the dominant fluid. In practice, two cases are considered: the radiation and the matter-dominated phases. Some simulations also discuss the transition between these two periods.

In both eras, it was found that the total energy density indeed tends towards a scaling law. More precisely, numerical simulations lead to [42]

$$\rho_{\infty}^{(\text{mat})} \simeq 28 \frac{U}{a^2 \eta^2} \quad \text{and} \quad \rho_{\infty}^{(\text{rad})} \simeq 40 \frac{U}{a^2 \eta^2},$$

hence confirming the energy-transfer mechanism.

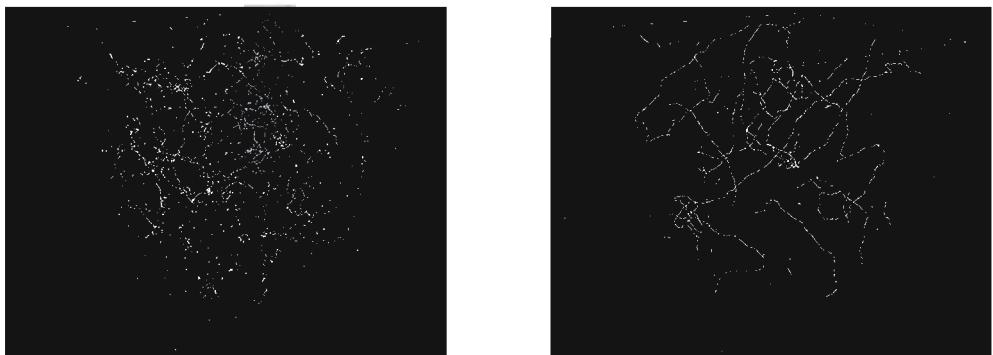


Fig. 11.13 Final state of the evolution of a cosmic-string network, on the right for a matter-dominated epoch, and on the left for a radiation dominated one. We observe the distribution of some infinite strings in the Hubble volume together with a large number of strings of any size. Figure produced and provided by C. Ringeval thanks to the original numerical simulation code by D. Bennett and F. Bouchet.

11.9.3.5 Cosmic microwave background angular power spectrum

Thanks to numerical simulations, it is possible to compute the deformation suffered by an electromagnetic plane wave passing through the string network due to the Kaiser-Stebbins effect. These strings simulations thus allow us to draw up temperature maps

¹⁰If the expansion is accelerated, i.e. if $w \leq -\frac{1}{3}$, it is the other term that quickly dominates, and the string contribution soon becomes negligible.

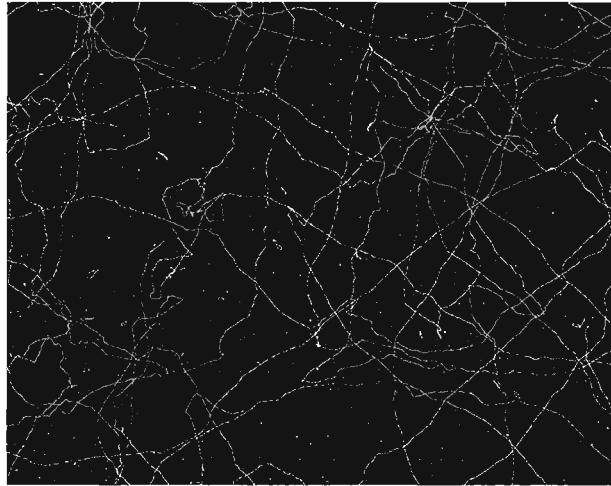


Fig. 11.14 Zoom in the final state of the evolution of a cosmic-string network in the matter-dominated era. We clearly see a large number of loops formed during the network evolution. It is these loops that allow for the evacuation of most of the energy contained in the strings and its dissipation in the form of gravitational radiation. Figure produced and provided by C. Ringeval thanks to the original numerical simulation code by D. Bennett and F. Bouchet.

of the cosmic microwave background, and, after averaging over a set of simulations, to calculate the cosmic microwave background angular power spectrum.

To start with, simulations make it possible to establish what is the value of the mean temperature fluctuation produced by the strings at very large angular scales (the plateau) in the cosmic microwave background. It is found to be

$$\left(\frac{\Delta T}{T} \right)_{\text{plateau}} \simeq 19 G_N U.$$

Since observations reveal that fluctuations are of the order of 10^{-5} , we obtain the upper bound $G_N \mu \lesssim 10^{-6}$, indicating a symmetry-breaking scale of the right order of magnitude for grand unification. Initially, it is this coincidence that stimulated an interest for the study of cosmic string networks.

The angular power spectrum of cosmic microwave background anisotropies has not yet been established with certainty for the moment, but some of its properties are already known. For instance, we already know that since the perturbations produced by a string network are not coherent, the spectrum does not present the standard acoustic peaks structure (Chapter 6). The different simulations realized so far indicate, but with important error bars, that the angular power spectrum starts with a small slope plateau at large angles (small ℓ), and then slowly raises to reach a maximum at around $\ell \simeq 350$ before going down very quickly thereafter.

Despite the absence of a precise theoretical curve for the angular power spectrum, available experimental data allow us to conclude that they cannot be responsible for all the fluctuations. Nevertheless, it still remains possible that a fraction is due to strings.

In this case, data indicate that the part of primordial fluctuations due to strings must be lower than around 10% [43]. As will be seen later, this already sets very severe constraints on some theories.

11.9.4 Other consequences of cosmic strings

11.9.4.1 Non-Abelian strings

Much more complicated strings can be produced during an arbitrary phase transition. In particular, since the initial group is most of the time non-Abelian, it is not always possible to factor the breaking scheme into that of a U(1) subgroup times a non-defect producing group. In this case, the breaking leading to strings is then that of a non-Abelian group, and the resulting strings are accordingly called non-Abelian.

More precisely, strings that can be produced according to the scheme (11.12) are classified according to the first homotopy group of \mathcal{G}/\mathcal{H} , and thus, from (11.41), $\pi_0(\mathcal{H})$. For a discrete resulting group \mathcal{H} , one has $\pi_0(\mathcal{H}) \sim \mathcal{H}$. If \mathcal{H} is non Abelian, the string will be said to be non-Abelian.

In the non-Abelian case, the generalization of the rotation of the Higgs field around the centre of a string, but far from it, is written as

$$\phi(\theta) = U(\theta)\phi(0) = e^{iT^s\theta}\phi(0),$$

where T^s is one of the group generators. This expression is actually not gauge invariant, but there exists a more complicated form that gives essentially similar results. For our purposes, the above form is sufficient. In the case where there is only one generator, the string is Abelian and we are back to the previous studies. If there are several string generators, and that, moreover, they do not commute, then the reconnection described earlier does not occur: indeed, if two strings, with respective generators T^1 and T^2 say, meet, they will be connected by a third string, with generator proportional to the commutator $[T^1, T^2]$.

The evolution of such a network is very different from that of an Abelian network. In particular, the presence of new string segments linking the intersecting ones prevents the creation of loops, and thus suppresses the most efficient mechanism to reduce the energy contained in the network. We can thus a priori doubt the existence of a scaling regime for such a network. Two simulations have been made on this topic. The first [44] showing that a string network with $\mathcal{H} = S_3$, group of the permutations with three elements, seems to reach a scaling regime with many more strings than an Abelian network, and especially with a density strongly dependent on the initial conditions. The second simulation [45] indicates that a global string network with $\mathcal{H} = Z_3$ seems to slow down the formation of a so-called ‘frustrated’ network. In such a network, since each intersection introduces a new string, the initial network kinetic energy is slowly transferred in these new strings and the network ends up becoming solid [46] with only the dynamics being that of the expansion. Such a network behaves as a dark-matter fluid with equation of state $w = -\frac{1}{3}$. The difference between these two simulations could arise from the fact that global strings lose much more energy than local strings since the first ones do it via Goldstone radiation, which is very efficient, while the others rest on gravitational radiation, which is a much slower process.

11.9.4.2 Gravitational radiation

When a local string loop is in motion, it can reach an equilibrium configuration due to its angular momentum allowing it to resist the contraction force from the tension. Such configurations would have an infinite lifetime if gravitational radiation were absent.

For a loop of length L , mass $M = UL$ and thus quadrupole momentum $D \sim ML^2 \sim UL^3$, the quadrupole formula (see Section 1.132 of Chapter 1) for gravitational radiation gives the total rate of energy emission

$$\frac{dE}{dt} \sim G_N \left(\frac{d^3 D}{dt^3} \right)^2 \sim G_N M^2 L^4 \omega^6,$$

where ω is a characteristic frequency ($\omega \sim L^{-1}$). We thus find

$$\frac{dE}{dt} \sim \Gamma G_N U,$$

with Γ a numerical factor depending only on the form of the loop and on its trajectory. Numerical simulations and known analytical solutions indicate that it is a good approximation to set $\Gamma \sim 100$.

The lifetime of a grand unification loop is thus

$$\tau_{\text{loop}} = M \left(\frac{dE}{dt} \right)^{-1} \sim \frac{L}{\Gamma G_N U} \sim 10^4 L, \quad (11.54)$$

showing that strings take a long time to decay via gravitational radiation.

The gravitational radiation emitted by a string network is obtained after integration over the whole ensemble of loops. By doing so, we obtain a stochastic gravity wave background that can be compared to that expected in different models of inflation [47] (Fig. 11.15) on which a non-Gaussian component is superposed, induced from the ‘explosive’ gravitational radiation at the strings ‘cusps’. This last component would be detectable by LIGO, LISA or pulsars if $G_N U > 10^{-13}$. For now, insofar as no gravitational waves have yet been detected, the best constraint on this gravity wave background produced by a set of strings arises from the stability of pulsars. When gravitational waves pass between these pulsars and the Earth, they perturb the metric and thus produce fluctuations in the arrival times of the pulses. Since this is a continuous signal in time, it is possible to improve the measured precision simply by integrating over the measured time. This is why it is to date the best known constraint on the strings energy scale. It leads to

$$G_N U \lesssim 10^{-5}. \quad (11.55)$$

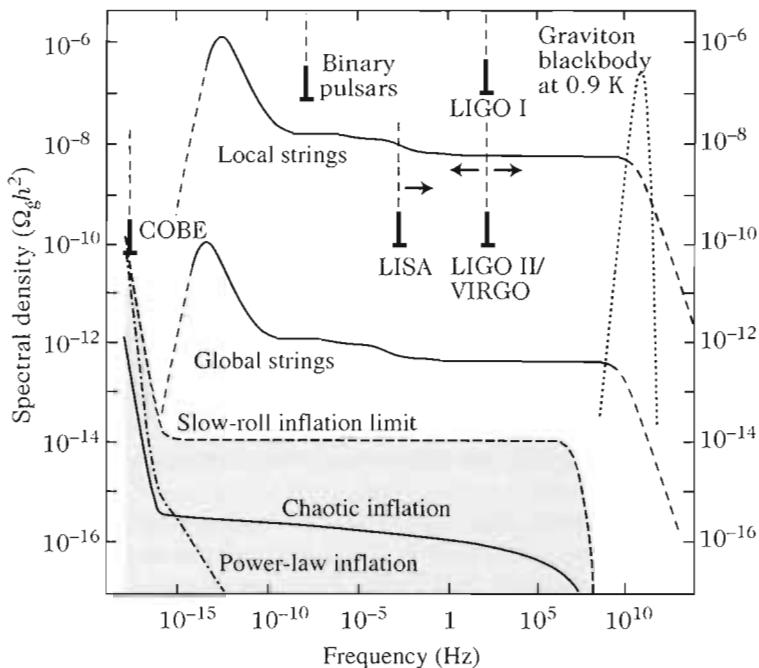


Fig. 11.15 Gravitational radiation spectrum of a string network [47]: both dotted curves show the contribution of the strings that decay during the radiation- and matter-dominated periods. For comparison, the spectrum of the gravitational radiation background at 0.9 K, analogous to the cosmic microwave background, is indicated, as well as the contributions due to inflation and to a strongly first-order electroweak transition phase. We have also placed the constraint from pulsar stability as well as the expected sensitivity of gravitational waves detection experiments. From Ref. [48].

References

- [1] H. E. STANLEY, *Introduction to phase transitions and critical phenomena*, Oxford University Press, 1971; N. GOLDENFELD, *Lectures on phase transitions and the renormalization group*, Addison-Wesley, 1992. The more recent and up-to-date following reference could also be of interest to French speaking readers, namely M. LAGUËS and A. LESNE, *Invariances d'échelles – Des changements d'état à la turbulence*, Belin, 2003 (in French, not translated). A. LESNE, *Renormalization methods*, Wiley, 1998 is slightly more specialized.
- [2] An informal introduction to some topological defects and their cosmological relevance in cosmology can be found in A. GANGUI, *Superconducting Cosmic Strings American Scientist*, Vol 88, N° 2, pages 254–263, May–June issue (2000). [astro-ph/0005186]. This article, heavily based on the popular book entitled *Des défauts dans l'Univers* by P. PETER and A. GANGUI, CNRS Editions, 2003 (in French, not translated), also contains many related references.
- [3] D. BAILIN and A. LOVE, *Introduction to gauge field theory*, Intitute of Physics Publishing, 1986, Chapter 17.
- [4] M. NAKAHARA, *Geometry, topology and physics*, Institute of Physics Publishing, 1990.
- [5] C. RINGEVAL, P. PETER and J.-P. UZAN, ‘Localization of massive fermions on the brane’, *Phys. Rev. D* **65**, 044016, 2002, and references therein.
- [6] C. RINGEVAL, P. PETER and J.-P. UZAN, ‘Stability of six-dimensional hyperstring braneworlds’, *Phys. Rev. D* **71**, 104018, 2005, and references therein.
- [7] B. DIU, C. GUTHMANN, D. LEDERER and B. ROULET, *Physique statistique*, Hermann, 2001; M. LE BELLAC, *Quantum and statistical field theory*, Oxford University Press, 1992.
- [8] M. B. HINDMARSH and T. W. B. KIBBLE, ‘Cosmic strings’, *Rep. Prog. Phys.* **58**, 477, 1995.
- [9] H. B. NIELSEN and P. OLESEN, ‘Vortex-line models for dual strings’, *Nucl. Phys. B* **61**, 45, 1973.
- [10] Y. NAMBU, *Symmetries and quark model*, R. Chand (ed.), Gordon & Breach, 1970; T. GOTO, ‘Relativistic Quantum Mechanics of One-Dimensional Mechanical Continuum and Subsidiary Condition of Dual Resonance Model’, *Prog. Theor. Phys.* **46**, 1560, 1971.
- [11] E. WITTEN, ‘Superconducting strings’, *Nucl. Phys. B* **249**, 557, 1985.
- [12] P. PETER, ‘Superconducting cosmic string: Equation of state for space-like and time-like currents in the neutral limit’, *Phys. Rev. D* **45**, 1091, 1992; ‘Influence of the electric coupling strength in current-carrying cosmic strings’, *Phys. Rev. D* **46**, 3335, 1992.
- [13] B. CARTER and P. PETER, ‘Supersonic string model for Witten vortices’, *Phys.*

- Rev. D* **52**, R1744, 1995; B. CARTER, P. PETER and A. GANGUI, ‘Avoidance of collapse by circular current-carrying cosmic-string loops’, *Phys. Rev. D* **55**, 4647, 1997.
- [14] R. L. DAVIS and E. P. S. SHELLARD, ‘Cosmic vortons’, *Nucl. Phys. B* **323**, 209, 1989.
- [15] R. BRANDENBERGER *et al.*, ‘Cosmic vortons and particle-physics constraints’, *Phys. Rev. D* **54**, 6059, 1996.
- [16] A. VILENKO and E. P. S. SHELLARD, *Cosmic strings and other topological defects*, Cambridge University Press, 2000.
- [17] G. T. HOOFT, ‘Magnetic monopoles in unified gauge theories’, *Nucl. Phys. B* **79**, 276, 1974; A. M. POLYAKOV, ‘Particle spectrum in quantum field theory’, *JETP Lett.* **20**, 194, 1974.
- [18] E. HUGUET and P. PETER, ‘Bound states in monopoles: sources for UHECR?’, *Astropart. Phys.* **12**, 277, 2000.
- [19] E. B. BOGOMOL’NYI, ‘The stability of classical solutions’, *Sov. J. Nucl. Phys.* **24**, 449, 1976; M. K. PRASAD and C. M. SOMMERFIELD, ‘Exact classical solution for the ’t Hooft monopole and the Julia-Zee dyon’, *Phys. Rev. Lett.* **35**, 760, 1975.
- [20] C. NASH and S. SEN, *Topology and geometry for physicists*, Academic Press, 1987.
- [21] D. STAUFFER and A. AHARONY, *Introduction to percolation theory*, Taylor and Francis, 1994; D. STAUFFER, ‘Scaling theory of percolation clusters’, *Phys. Rep.* **54**, 1, 1979.
- [22] P. PETER, ‘Surface current-carrying domain walls’, *J. Phys. A* **29**, 5125, 1996.
- [23] A. E. EVERETT, ‘Observational consequences of a ‘domain’ structure of the Universe’, *Phys. Rev. D* **10**, 3161, (1974); Ya B. ZEL’DOVICH *et al.*, *Sov. Phys. JETP* **40**, 1, 1975.
- [24] C. T. HILL, ‘Monopolonium’, *Nucl. Phys. B* **224**, 469, 1983.
- [25] D. A. DICUS, D. N. PAGE and V. L. TEPLITZ, ‘Two- and three-body contributions to cosmological monopole annihilation’, *Phys. Rev. D* **26**, 1306, 1982.
- [26] J. P. PRESKILL, ‘Cosmological production of superheavy monopoles’, *Phys. Rev. Lett.* **43**, 1365, 1979.
- [27] M. B. EINHORN, D. L. STEIN and D. TOUSSAINT, ‘Are grand unified theories compatible with standard cosmology?’, *Phys. Rev. D* **21**, 3295, 1980.
- [28] G. DVALI, A. MELFO and G. SENJANOVIC, ‘Is there a monopole problem?’, *Phys. Rev. Lett.* **75**, 4559, 1995.
- [29] M. S. TURNER, ‘Thermal production of superheavy magnetic monopoles in the early Universe’, *Phys. Lett. B* **115**, 95, 1982.
- [30] G. DVALI, H. LIU and T. VACHASPATI, ‘Sweeping away the monopole problem’, *Phys. Rev. Lett.* **80**, 2281, 1998.
- [31] P. LANGACKER and S. -Y. PI, ‘Magnetic monopoles in grand unified theories’, *Phys. Rev. Lett.* **45**, 1, 1980.
- [32] R. JEANNEROT, J. ROCHER and M. SAKELLARIADOU, ‘How generic is cosmic-string formation in supersymmetric grand unified theories?’, *Phys. Rev. D* **68**, 103514, 2003.

- [33] P. PETER, ‘Comments on some metric properties of cosmic strings having a non-degenerate stress-energy tensor’, *Class. Quant. Grav.* **11**, 131, 1994.
- [34] M. SAZHIN et al., ‘CSL-1: chance projection effect or serendipitous discovery of a gravitational lens induced by a cosmic string’, *Month. Not. R. Astron. Soc.* **343**, 353, 2003; ‘Lens candidates in the Capodimonte Deep Field in vicinity of the CSL-1 object’, [astro-ph/0406516]; R. SCHILD et al., ‘Anomalous fluctuations in observations of Q0957+561 A,B: smoking gun of a cosmic string?’, *Astron. Astrophys.* **422**, 477, 2004; E. AGOL, C. J. HOGAN and R. M. PLOTKIN, ‘Hubble Imaging Excludes Cosmic-String Lens’, *Phys. Rev. D* **73**, 087302, 2006; M. SAZHIN et al., ‘Gravitational lensing by cosmic strings: What we learn from the CSL-1 case’, *Mon. Not. R. Astron. Soc.* **376**, 1731, 2007.
- [35] J.-P. UZAN and F. BERNARDEAU, ‘Cosmic lens phenomenology: general properties of distortion field’, *Phys. Rev. D* **63**, 023004, 2000; F. BERNARDEAU and J.-P. UZAN, ‘Cosmic lens phenomenology: model of Poisson energy distribution’, *Phys. Rev. D* **63**, 023005, 2000.
- [36] N. KAISER and A. STEBBINS, ‘Microwave anisotropy due to cosmic strings’, *Nature* **310**, 391, 1984.
- [37] T. VACHASPATI and A. VILENKO, ‘Formation and evolution of cosmic strings’, *Phys. Rev. D* **30**, 2036, 1984.
- [38] F. R. BOUCHET, D. P. BENNETT and A. STEBBINS, ‘Patterns of the cosmic microwave background from evolving string networks’, *Nature* **335**, 410, 1988.
- [39] T. W. B. KIBBLE, ‘Topology of cosmic domains and strings’, *J. Phys. A* **9**, 1387, 1976 and ‘Some implications of a cosmological phase transition’, *Phys. Rep.* **67**, 183, 1980.
- [40] E. COPELAND, T. W. B. KIBBLE and D. AUSTIN, ‘Scaling solutions in cosmic-string networks’, *Phys. Rev. D* **45**, R1000, 1992; D. AUSTIN, E. COPELAND and T. W. B. KIBBLE, ‘Evolution of cosmic string configurations’, *Phys. Rev. D* **48**, 5594, 1993.
- [41] C. J. A. P. MARTINS and E. P. S. SHELLARD, ‘Quantitative string evolution’, *Phys. Rev. D* **54**, 2535, 1996; ‘Extending the velocity-dependent one-scale string evolution model’, *Phys. Rev. D* **65**, 043514, 2002.
- [42] C. RINGEVAL, *Fermionic currents flowing along extended objects*, PhD thesis, Pierre et Marie Curie University, Paris [hep-ph/0211126].
- [43] F. R. BOUCHET, P. PETER, A. RIAZUELO and M. SAKELLARIADOU, ‘Is there evidence for topological defects in the BOOMERanG data?’, *Phys. Rev. D* **65**, 021301, 2002; L. POGOSIAN, M. WYMAN and I. WASSERMAN, ‘Observational constraints on cosmic strings: Bayesian analysis in a three dimensional parameter space’, *JCAP* **09**, 008, 2004.
- [44] P. McGRAW, ‘Evolution of a non-Abelian cosmic-string network’, *Phys. Rev. D* **57**, 3317, 1998.
- [45] D. SPERGEL and U.-L. PEN, ‘Cosmology in a string-dominated Universe’, *Astrophys. J. Lett.* **491**, L67, 1997.
- [46] M. BUCHER and D. N. SPERGEL, ‘Is the dark matter a solid?’, *Phys. Rev. D* **60**, 043505, 1999.

- [47] B. ALLEN, The stochastic gravity-wave background: sources and detection, Proceedings of the ‘Les Houches School on Astrophysical Sources of Gravitational Waves’, J.-A. Marck and J.-P. Lasota (eds.), Cambridge University Press, 1996, [gr-qc/9604033]
- [48] R. A. BATTYE, R. R. CALDWELL and E. P. S. SHELLARD, *Gravitational waves from cosmic strings*, [astro-ph/9706013], Proceedings of ‘Topological defects in cosmology’, M. Signore and F. Melchiorri (eds.), World Scientific, 1998.

12

Extensions of the standard cosmological framework

This chapter illustrates various extensions of the standard cosmological framework considering the theoretical developments presented in Chapter 10. Without being exhaustive we give, in Section 12.1, an overview of the contribution of supersymmetry and supergravity to the construction of inflationary models. As for the recent Universe, we explore the question of the origin of dark energy, Section 12.2, which is today at the heart of many considerations. Finally, we mention the question of the constancy of the fundamental constants, Section 12.3, and the topology of the Universe, Section 12.4, two subjects that have experienced significant progress recently.

12.1 Construction of inflationary models

The inflationary paradigm is now a cornerstone of the standard cosmological model (see Chapter 8 for the arguments). However, to set the physics of inflation on a more solid basis, one should embed it into the general framework of field theory that describes particles and their interactions as we know them. We should thus be able to formulate theories containing the standard model (Chapter 2) in some low-energy limit and containing scalar fields with sufficiently flat potentials to give rise to a phase of inflation.

Grand unification (Chapter 10) and string theory (Chapter 13), and in fact, any consistent high-energy extension of the standard model, up to now, demands that supersymmetry be implemented. This generic property turns out to provide the necessary scalar fields, and leads to essentially two main families of models that might tackle the inflation issue. These models are based on the F or D term of supersymmetry, or even supergravity. Here, we shall present the basic requirements for building such models, with simple examples, and recommend the reviews [1, 2] in which more explicit (and realistic) models have been analyzed.

12.1.1 A consistent model: supergravity

The theory from which the inflationary model will be constructed describes a large number of particles, their masses, and their interaction terms. In order for this theory to be physically realistic, it should contain all the information on the standard model described in Chapter 2. Equivalently, one should be able to embed this model into an extended framework that contains a sector capable of giving rise to a phase of inflation.

In order for the inflationary phase to be related with observational data, its characteristic energy scale ought to be of the order of that of grand unification (Chapter 8). Thus, we shall take for granted that supersymmetric grand unified theories are good candidates for the construction of a realistic model. As discussed above, the choice of supersymmetry can be justified by two arguments. First, no realistic non-supersymmetric theory containing the standard model, while at the same time being compatible with all the (particle-physics) experiments, has yet been found. Moreover, and this is the second reason, string theory must contain this symmetry in order to be stable.

Two cases can then arise. Either we want the model to be renormalizable (or finite, which could be arguably even better), or we merely want to investigate the consequences of an effective, non-renormalizable model. The former case is somewhat favoured by many for the following reasons: a renormalizable theory can only contain a finite number of parameters that must ultimately be determined experimentally. Besides, the standard model itself is renormalizable, and so is its global supersymmetric extension.

An advantage of a renormalizable theory is that it can be valid at all energy scales, or, more likely, as a low-energy approximation of a more general theory, valid below a given characteristic scale, E_c say. This a priori unknown theory can be expanded at low energy: in the limit $E_c \rightarrow \infty$, we recover a renormalizable theory, i.e. one that does not depend on E_c . The next-order terms are obviously undetermined, but can be constructed in a phenomenological way using dimensional arguments. We then find terms for which the effective coupling constants behave as E_c^{-n} , with the power $n = 2, 4, \dots$. These are usually non-renormalizable, but that need not worry us for we are dealing, after all, with an effective theory. In practice, one often fixes E_c to be of the order of either the Planck mass M_p or the grand unification scale E_{GUT} . Around and beyond this typical scale, corrections due to the non-renormalizable terms, i.e. from the full theory, may become important and must be evaluated.

Now we want to consider such models in order to apply them to a cosmological case, namely the inflation epoch. Therefore, we must explicitly consider a coupling to gravity: the limit $M_p \rightarrow \infty$ can in no way be taken, as it is meaningless here. All possible extra terms coming in particular from supergravity should thus be taken into account, again including the non-renormalizable terms.

Among the numerous fields present in supersymmetric grand unification theories,¹ one should single out the scalar fields that are capable of dominating the matter content of the Universe. These scalar fields appear in the Lagrangian of the theory, in particular in the so-called F and D terms. We shall therefore look at these terms in turn later.

12.1.1.1 ‘Flat potential’ constraints

The fact that the theory should be realistic in the particle-physics sense, although clearly necessary, is not sufficient. In particular, the potential of the scalar field that will dominate the dynamics of the Universe should also have the right properties to

¹The acronym SUSY-GUT, or simply SGUT, is sometime used in the literature; we shall, however, avoid this practice in the following.

give rise to a sufficiently long inflationary phase (recall the observational constraints derived in Chapter 8). In practice, this potential should be flat enough, at least in some directions, so that the slow-roll conditions² can be implemented.

However, a general problem arises. A scalar field whose dynamics is driven by a flat potential also has, as a consequence of flatness, a very small mass. If this light scalar field is coupled to the other fields of the theory, the quantum corrections due to the interactions will induce large modifications to this potential, and might thus drastically change the mass of the scalar field. Once these corrections are taken into account, such a theory ends up, in general, being no longer a serious candidate since the scalar field becomes too massive to produce super-Hubble correlations.

12.1.1.2 Coleman–Weinberg correction

Once the potentials of all the fields of the theory have been chosen, one can compute the quantum corrections due to the interactions between the different fields of the model. The scalar potentials, in addition to the D and F terms, then obtain an additional term ΔV_{loop} , originally calculated by Coleman and Weinberg [3], which reads

$$\Delta V_{\text{loop}} = \frac{1}{64\pi^2} \sum_i (-1)^{N_F} m_i^4 \ln \left(\frac{m_i}{\Lambda} \right)^2, \quad (12.1)$$

where the sum is performed over all the helicity states i . Here, N_F is the fermionic number (Section 10.3 of Chapter 10) and m_i the masses of the states of helicity i . Equation (12.1) was obtained by renormalizing at a cutoff energy scale Λ . It is absolutely mandatory to take this correction into account in order to calculate correctly the cosmological predictions of the supersymmetric models of inflation.

To illustrate the effect of this correction, let us consider a potential that can be expanded along one of its flat directions as

$$V = V_0 + \frac{1}{2} m^2(\phi) \phi^2 + \dots, \quad (12.2)$$

where the dots correspond to non-renormalizable terms. Choosing the Planck scale as a cutoff, i.e. setting $\Lambda = M_P$, the one-loop correction generically takes the form

$$m^2(\phi) = m^2 + c \mu^2 \ln \left(\frac{\phi}{M_P} \right), \quad (12.3)$$

where μ is a mass scale and $c \ll 1$ is a pure constant. It is then easy to convince oneself that if $\mu \sim m$, the effective potential has an extremum for $\phi \sim M_P \exp(-1/c)$. This extremum is a minimum if $m(M_P)^2 > 0$. Upon plotting these potentials, we find that they correspond to potentials similar to those shown in Fig. 8.2, leading to the ‘old’ or ‘new’ inflation models of Chapter 8.

²We could imagine implementing inflation in a consistent way without, however, satisfying these conditions, but it should then be demonstrated explicitly for each case. On the other hand, if the slow-roll hypothesis is valid, we are ensured that the theory will be able to satisfy the observational constraints.

Let us, however, stress a possible shortcoming concerning the above conclusion. The loop correction terms are calculated from the field theory in flat space-time, i.e. in a Minkowski space, whereas during inflation the space-time structure is quasi-de Sitter. Rigorously, one should compute these corrections in a quantum field theory in the curved space-time formalism (see Ref. [4] for an explicit example computing the mass renormalization in a de Sitter space-time). In most cases, this does not affect the conclusion, but one should check that it is indeed so before comparing with any data.

12.1.1.3 Strategies

The required constraints on the potential at tree level to obtain a slow-roll inflationary phase are described in Chapter 8. Here, we shall discuss how to fulfill those in the present context of realistic theories.

Within the framework of global supersymmetry, supergravity corrections are not small during inflation. They typically imply that $m^2 \gtrsim V_0/M_p^2$ for every scalar field and in particular for the inflaton. For the latter, $V_0 \sim H^2 M_p^2$ in the slow-roll regime so that its mass becomes greater than the Hubble constant during inflation, i.e. $m \gtrsim H$; it is thus too ‘heavy’ to produce super-Hubble correlations. Various strategies have been proposed to solve this problem and construct a viable model in this context,

- the potential is dominated by an F -term but the mass of the inflaton is suppressed because the Kähler potential, K , and the superpotential, W , have very specific forms;
- the potential is dominated by a generic F -term but the mass of the inflaton is suppressed due to a cancellation between the different terms;
- the potential is dominated by a generic F -term at the Planck scale M_p but the mass of the inflaton is suppressed at energy scales where inflation occurred, due to a scale dependence;
- the potential is dominated by a D -term;
- the potential is dominated by a generic F -term but the kinetic term of the inflaton becomes singular in the regime where inflation occurred so that once the kinetic term has its canonical form, the potential becomes sufficiently flat even if it was not originally.

In what follows, for the sake of simplicity, we illustrate the two possible situations (F or D) with a theory containing three chiral multiplets, Φ_0 , Φ_+ and Φ_- . As for the potentials, they will depend on the particular case considered and we will choose them as appropriate.

12.1.2 F -term inflation

The first possibility is thus that for which the potential is dominated by the F -term, which amounts to saying that the contributions of the D -terms have only a limited, and in most cases negligible, influence on the dynamics of the Universe. Notice that these D -terms can nonetheless play an important role in the sense that they can determine the inflaton trajectory.

12.1.2.1 Effective potential

Let us consider the case of a renormalizable superpotential

$$W_F = \lambda \Phi_0 (\Phi_+ \Phi_- - M^2), \quad (12.4)$$

where λ is a coupling constant and M a mass assumed to be very small compared to the Planck mass, $M \ll M_P$, in order for supergravity corrections to be negligible. In this case, (10.118) allows us to derive the potential

$$V_F = \lambda^2 \left[|\phi_+ \phi_- - M^2|^2 + |\phi_0|^2 \left(|\phi_+|^2 + |\phi_-|^2 \right) \right], \quad (12.5)$$

whose global minimum is localized at $\phi_0 = 0$ and $\phi_+ = \phi_- = M$, with $V_F = 0$, and thus satisfying supersymmetry.

Identifying ϕ_0 with the inflaton and assuming chaotic initial conditions, then there should exist regions of space for which $\phi_0 > M$. Consider one such region: the potential has a valley of local minima for $\phi_+ = \phi_- = 0$, and there is thus a constant term in the potential $V_F \supset \lambda^2 M^4$. Since $V_F > 0$, supersymmetry is broken during the inflation phase, and the potential does not depend on the inflaton ϕ_0 .

Now, to evaluate the one-loop corrections given by (12.1), we need to know exactly the particle spectrum of the theory. Since supersymmetry is broken, we have a Dirac fermion of mass $\lambda |\phi_0|$ and two scalar fields of masses $\lambda \sqrt{|\phi_0|^2 \pm M^2}$. Equation (12.1) then implies that

$$V_F^{(\text{eff})} = \lambda^2 M^4 \left(1 + \frac{\lambda^2 N}{32\pi^2} \Delta \right), \quad (12.6)$$

with

$$\Delta = 2 \ln \left(\frac{\lambda^2 M^2}{\Lambda^2} z \right) + (z+1)^2 \ln(1+z^{-1}) + (z-1)^2 \ln(1-z^{-1}), \quad (12.7)$$

where N is the dimension of the representation in which the superfields Φ_\pm are, and we have set $z = |\phi_0|^2 / M^2$.

The potential (12.6) serves as a starting point for a scenario of hybrid inflation: as soon as the inflaton ϕ_0 reaches the critical value M , it becomes energetically favourable for ϕ_\pm to take non-vanishing values, evolving dynamically towards the supersymmetric global minimum. The parameters of this model are then constrained by observations, in particular that of the cosmic microwave background. When implemented for the grand unification group $\text{SO}(10)$ [5], we find that we should have

$$M \lesssim 2 \times 10^{15} \text{ GeV} \quad \text{and} \quad \lambda \lesssim 7 \times 10^{-7},$$

where we have also taken into account the fact that cosmic strings, whose contribution to the anisotropies [6] must be lower than around 10%, are generically produced in this kind of models [7] (and this depends only slightly on the choice of the gauge group).

Many other models can similarly be constructed. A compilation including a description of many such proposals with their respective constraints can be found, for instance, in Ref. [2].

12.1.2.2 The η problem

The previous scenario suffers, however, from a problem if one wishes to include supergravity corrections. This stems from the fact that if one directly applies (10.142), one finds a potential containing a global pre-factor of the form $V \sim \exp(8\pi G_N \phi_0^2) [\dots]$, so that the slow-roll parameter η_V , defined by (8.49), turns out to be of order unity. Indeed, differentiating the exponential of the Kähler potential, we find that $\eta_V = 1 + \dots$, and the additional terms have a priori no reason to cancel the first one, unless of course if the Kähler potential and the superpotential are artificially tuned with this intention.

There exist many solutions to this problem. The simplest is to not consider corrections from supergravity at all! This is possible as in this category of models, the inflaton never takes values large compared to the Planck mass (i.e. $\phi < M_P$). Considering a Kähler potential of the form (10.141) then the first mass term of the inflaton simplifies and there only remains higher-order terms, here behaving as $|\phi_0/M_P|^4 \ll 1$.

One can also impose special symmetries on the theory that avoid the appearance of such impeding terms. Moreover, one can also construct perfectly consistent models of inflation for which $\eta_V \sim 1$ is not problematic. This is possible provided $\varepsilon \ll 1$, and should be examined on a case by case basis.

The last solution is, of course, to turn to different kinds of models, where this problem is absent. As this is, for instance, the case when the D -term dominates the total potential, we now turn to this second category.

12.1.3 D -term inflation

Let us now consider a theory invariant under the transformations of the U(1) group, so that the Fayet–Iliopoulos D -term is important. We then choose the superpotential

$$W_D = \lambda \Phi_0 \Phi_+ \Phi_-, \quad (12.8)$$

which is essentially the same as for the F -term case but with a mass parameter $M = 0$. In W_D , the superfields have charges under U(1) given by $Q(\Phi_0) = 0$ and $Q(\Phi_{\pm}) = \pm 1$. The inflaton will then be identified with the radial part of a field of vanishing charge.

12.1.3.1 Global supersymmetry

The superpotential (12.8) leads to a potential arising from the F -term, and the theory also contains a D -term, given by (10.125), so that the total potential is the sum of both terms,

$$V_D = \lambda^2 \left(|\phi_+ \phi_-|^2 + |\phi_0 \phi_+|^2 + |\phi_0 \phi_-|^2 \right) + \frac{g^2}{2} \left(\xi + |\phi_+|^2 - |\phi_-|^2 \right)^2, \quad (12.9)$$

where g is the coupling constant of the U(1) group. For definiteness, we also assume that $\xi > 0$. The F part of the potential is then positive definite and we therefore only have one minimum localized at zero, when at least two of the fields (ϕ_0 and ϕ_+ or ϕ_-) vanish. The Fayet–Iliopoulos term, on the other hand allows us to break the U(1) symmetry in the sense that the potential has a non-vanishing minimum at $\phi_- = \sqrt{\xi}$. There therefore exists a global supersymmetric minimum ($V_D = 0$) for

$$\phi_0 = \phi_+ = 0, \quad \phi_- = \sqrt{\xi}.$$

If the inflaton ϕ_0 takes a value larger than the critical value ϕ_c given by

$$\phi_c = \left(\frac{g^2 \xi}{2\lambda^2} \right)^{1/2},$$

then we recover a local minimum at $\phi_+ = \phi_- = 0$. So for $\phi > \phi_c$ and $\phi_+ = \phi_- = 0$, the tree-level potential is flat in the direction of ϕ_0 where it has a value independent of ϕ_0 , namely

$$V_D = \frac{1}{2} g^2 \xi^2.$$

The fields ϕ_+ and ϕ_- are massive and the mass spectrum consists of a Dirac fermion of mass $\lambda |\phi_0|$ and two scalar fields of mass $m_\pm^2 = \lambda^2 |\phi_0|^2 \pm g^2 \xi$. This simple theory allows for a realization of a model of ‘hybrid’ inflation (see Fig. 8.7).

Having determined the mass spectrum, one now needs to evaluate the radiative correction [8] it induces from (12.1). We then obtain the effective potential

$$V_D^{(\text{1loop})} = \frac{1}{2} g^2 \xi^2 \left[1 + \frac{g^2}{16\pi^2} \ln \left(\frac{\phi_0^2}{\Lambda^2} \right) \right]. \quad (12.10)$$

In this model, inflation in the slow-roll regime stops when either ϕ_0 reaches ϕ_c or when $\eta_V \sim 1$, which occurs for $\phi_* \sim \sqrt{g/8\pi^2 M_p}$, depending on the value of the parameters. If inflation ends with the end of the slow-roll regime, the field ϕ starts oscillating until its amplitude becomes smaller than ϕ_c .

12.1.3.2 Supergravity corrections

Unlike the case of F -term inflation, supergravity corrections can be quite important in D -term inflation. However, it turns out that they do not prevent the model from being compatible with the data. We find that the boson and fermion masses, which depend on $\exp(|\phi_0|^2/M_p^2)$, do modify the effective potential (12.1), rendering the slow-roll calculations slightly more complicated to perform. Observational constraints then impose [5]

$$g \lesssim 10^{-2}, \quad \lambda \lesssim 10^{-5} \quad \text{and} \quad \sqrt{\xi} \lesssim 10^{15} \text{ GeV}.$$

Just as for the case of F -term inflation, many models can be constructed. Reference [2] again offers a compilation and description of many such models. Note in particular the existence of models built in the framework of grand unification.

12.2 Cosmological constant and dark energy

Cosmological observations indicate that the expansion of the Universe has recently entered a phase of acceleration, so that $q_0 < 0$ (Chapter 4). This conclusion relies only on the fact that the Universe is described by a Friedmann–Lemaître space-time, that is that the Copernican principle holds. If gravity is well described by the theory of general relativity, then such an acceleration can only be explained if the matter content of the Universe is dominated by a fluid whose equation of state satisfies $w < -1/3$.

The first candidate that naturally comes to one's mind for such a component is the *cosmological constant*, which corresponds to $w = -1$.

This section studies the different aspects concerning the physical nature of this component, generically called *dark energy* and describes some models that have been suggested to explain it. For reviews, one can refer to Refs. [9–11].

12.2.1 The cosmological constant problem

The cosmological constant is a dimensioned parameter (length^{-2}) that appears in the Einstein equations (1.85). Depending on whether this constant (denoted by Λ_0 for the moment) is arbitrarily written on the left- or the right-hand side, it can be interpreted as a geometrical term or as a term associated to some matter with stress-energy tensor

$$T_{\mu\nu} = -\frac{\Lambda_0}{8\pi G}g_{\mu\nu}. \quad (12.11)$$

This would then correspond to a fluid with equation of state $w = -1$ and constant energy density

$$\rho_\Lambda = \frac{\Lambda_0}{8\pi G}. \quad (12.12)$$

From the perspective of general relativity this parameter is completely free and there is no theoretical argument allowing us to fix the value of this cosmological constant, or equivalently, the length scale $\ell_\Lambda = |\Lambda_0|^{-1/2}$. Cosmology still roughly imposes that

$$|\Lambda_0| \leq H_0^2 \iff \ell_\Lambda \leq H_0^{-1} \sim 10^{26} \text{ m} \sim 10^{41} \text{ GeV}^{-1}.$$

In itself this value does not create any problem, as long as we only consider classical physics. Notice, however, that it is disproportionately large compared to the natural scale fixed by the Planck length

$$\ell_\Lambda \gtrsim 10^{60} \ell_p \iff \frac{\Lambda_0}{M_p^2} \lesssim 10^{-120} \iff \rho_\Lambda \lesssim 10^{-120} M_p^4 \sim 10^{-47} \text{ GeV}^4, \quad (12.13)$$

when expressed in terms of energy density.

12.2.1.1 Vacuum energy density

Particle physics provides a new perspective to this question. The Lorentz invariance of the vacuum implies that its energy-momentum tensor must take the form [12] $\langle T_{\mu\nu}^{\text{vac}} \rangle = -\langle \rho \rangle g_{\mu\nu}$.

From a classical point of view, let us consider a scalar field with potential $V(\varphi)$. Its action takes the form (1.106) and its energy-momentum tensor is given by (1.107). Its configuration of lowest energy is obtained for $\partial_\mu \varphi = 0$. In this configuration

$$T_{\mu\nu} = -V(\varphi_0)g_{\mu\nu}, \quad (12.14)$$

where φ_0 is the value of the field that minimizes the potential. There is a priori no reason to have $V(\varphi_0) = 0$. In fact, the mechanism of spontaneous symmetry breaking (Chapters 2 and 11) tells us that this energy can change during the evolution of

the Universe if phase transitions occur. For instance, let us consider a Higgs kind of potential,

$$V(\varphi) = V_0 - \frac{1}{2}\mu^2\varphi^2 + \frac{1}{4}g\varphi^4. \quad (12.15)$$

The configuration $\varphi = 0$ is unstable and the true vacuum is established at $\varphi^2 = \mu^2/g$, at which value the energy density is

$$\langle\rho\rangle = V_0 - \frac{\mu^4}{4g}. \quad (12.16)$$

Now, if $V_0 = 0$ then the vacuum energy is negative. In the electroweak model, it is of the order of $\langle\rho\rangle \simeq -g(200 \text{ GeV})^4 \sim -10^6 \text{ GeV}^4$: that is already an embarrassing 53 orders of magnitude larger than what is required by cosmological observations! This situation can be avoided if $\rho_\Lambda + V_0 - \mu^4/g \sim 10^{-47} \text{ GeV}^4$, but then why would such a fine tuning occur? Reciprocally, such a cancellation would imply the existence of an enormous cosmological constant at temperatures larger than the electroweak scale.

From a quantum point of view, the vacuum energy receives a contribution

$$\langle\rho\rangle_{\text{vac}} = \int_0^{k_{\max}} \frac{d^3 k}{(2\pi)^3} \frac{1}{2} \sqrt{k^2 + m^2} \simeq \frac{k_{\max}^4}{16\pi^2}, \quad (12.17)$$

arising from the zero-point energy. Fixing the cutoff frequency, k_{\max} , to the electroweak scale or to the Planck scale, this contribution is then of the order of

$$\langle\rho\rangle_{\text{vac}}^{\text{EW}} \sim (200 \text{ GeV})^4, \quad \langle\rho\rangle_{\text{vac}}^{\text{Pl}} \sim (10^{18} \text{ GeV})^4, \quad (12.18)$$

which implies a disagreement of, respectively, 60 to 120 orders of magnitude, again!

The cosmological constant problem thus amounts to understanding why

$$|\rho_V| = |\rho_\Lambda + \langle\rho\rangle_{\text{vac}}| \lesssim 10^{-47} \text{ GeV}^4, \quad (12.19)$$

or equivalently,

$$|\Lambda| = |\Lambda_0 + 8\pi G \langle\rho\rangle_{\text{vac}}| \lesssim 10^{-120} M_{\text{P}}^2, \quad (12.20)$$

i.e. why ρ_V is so small *today* and not during the entire history of the Universe.

12.2.1.2 Contribution of supersymmetry

As long as it is not broken, supersymmetry implies that the vacuum energy must be strictly zero. Indeed, as seen in Chapter 10, the generators, Q_α , of the supersymmetry transformations satisfy the anticommutation relations $\{Q_\alpha, \bar{Q}_\beta\} = 2(\sigma_\mu)_{\alpha\beta} P^\mu$, where the σ_i are Pauli matrices, $\sigma_0 = 1$ and P^μ is the momentum operator. As long as supersymmetry is not broken, the vacuum $|0\rangle$ must satisfy

$$Q_\alpha |0\rangle = \bar{Q}_\beta |0\rangle = 0. \quad (12.21)$$

We deduce that $\langle 0|P^\mu|0\rangle = 0$, so that the vacuum energy, namely $\langle 0|P^0|0\rangle$, must strictly vanish.

One can recover this result from the potential of a chiral field, whose expression from the superpotential, W , is (Chapter 10)

$$V(\varphi, \varphi^*) = \sum_i \left| \frac{\partial W}{\partial \varphi^i} \right|^2, \quad (12.22)$$

ignoring the gauge degrees of freedom for the present argument. In the supersymmetric phase, W must be stationary in φ so that V , being positive definite, must have $V = 0$ as a minimum. Quantum effects then do not change this conclusion since the symmetry between bosons and fermions cancels exactly all the loop corrections. So $\langle \rho \rangle_{\text{vac}} = 0$.

Nevertheless, supersymmetry is broken today, which implies that $\langle \rho \rangle_{\text{vac}} > 0$. If it is broken at around 1 TeV, the vacuum energy density would be of the order of $(1 \text{ TeV})^4$ below this temperature, which is again a disagreement with observations of 60 orders of magnitude. Note that such a scenario is actually catastrophic as it would imply that $\rho_\Lambda \sim -(1 \text{ TeV})^4$ during the entire history of the Universe!

This argument is, however, based on a global supersymmetry, whereas the entire argument implying gravity must be made in a context of local supersymmetry, *supergravity*. Expression (12.22) is then replaced by

$$V(\varphi, \bar{\varphi}) = e^{8\pi G K} \left(D_i W K^{i\bar{j}} D_{\bar{j}} \bar{W} - 24\pi G |W|^2 \right) \quad (12.23)$$

where $D_i W = \partial W / \partial \varphi^i + 8\pi G \partial G / \partial \varphi^i$, $K(\varphi^i, \bar{\varphi}^i)$ being the Kähler potential and $K_{i\bar{j}}$ its derivative with respect to the two fields, i.e. $K_{i\bar{j}} = \partial K / \partial \varphi^i \partial \bar{\varphi}^j$. The condition that supersymmetry is not broken now becomes $D_i \bar{W} = 0$, giving a position in field space where the potential V is usually negative. We can still imagine a scenario in which supersymmetry is broken in such a way that the term in brackets vanishes, but, once again, a considerable amount of fine tuning is needed. We are thus back to the previous situation.

12.2.2 The nature of dark energy

Current observations allow us to conclude that the function $H(z)$ deduced from observations cannot be reproduced by a mixture of matter ($w = 0$) and radiation ($w = 1/3$). In particular, the recent acceleration of the Universe requires the introduction of new degrees of freedom in the cosmological models. We generically call *dark energy* any model explaining the recent phase of acceleration of the Universe, independently of its physical origin.

12.2.2.1 New matter or modification of general relativity?

Just as for the existence of dark matter, the conclusion that there should exist a form of dark energy relies on its gravitational effects on the expansion of the Universe, itself being observed only through the properties of baryonic matter and photons.

Two directions are then possible (similarly as for the construction of ‘dark matter’ versus ‘MOND’ described in Chapter 7):

- one can introduce a new kind of matter, not expected in the standard model of particle physics, while assuming that gravity at cosmological scales is still described by general relativity. This new matter must effectively behave as a fluid

with equation of state smaller than $-1/3$. We will describe in what follows several candidates, usually reducing to the introduction of a light scalar field.

- one can assume that gravity, i.e. *the long-range force that cannot be screened*, is not described by general relativity on cosmological scales. As summarized in Chapter 1, modifications of general relativity are extremely constrained in particular in the Solar System. Any construction in this direction must first check that such a modification of gravity agrees with these constraints. Among the various models, let us mention either those introducing a non-minimally and/or non-universally coupled scalar field or the ones involving massive gravitons.

In both cases, one cannot avoid introducing new matter fields (i.e. new physical degrees of freedom) beyond that of the standard model of particle physics and the graviton. In this sense, there must exist some dark component, whether or not it is responsible for a new long-range interaction.

Figure 12.1 presents four universality classes that can encompass the models that have been proposed so far as candidates for the dark-sector physics.

To finish, let us emphasize a third possibility. The conclusion that $q_0 < 0$ relies on the fact that our Universe is well described by a Friedmann–Lemaître space-time, that is on the Copernican principle. Questioning this principle may offer an explanation of dark energy without introducing new physics, but only by considering new cosmological solutions to describe our Universe. We shall also discuss this possibility later.

Note also that the pure (and simplest of all) Λ CDM model is, at the moment, compatible with all observations. We shall thus consider it as our reference model.

12.2.2.2 Possible tests for a physics of the dark sector

The dark sector plays a more and more significant role in cosmological models, and so it is important to go beyond the phenomenology of a simple self-interacting scalar field. The reconstruction of the equation of state, or equivalently of $H(z)$, is an important, but far from sufficient, task since, as we shall show, we will be able to construct many models leading to the same function $H(z)$. It is thus important to go beyond the observation of this function (obtained, for instance, from type Ia supernovæ).

In particular, the four universality classes of models of Fig. 12.1 have specific signatures that we can try to detect. Recalling that the theory of general relativity was based on the Einstein equivalence principle, one way to sharpen the tests of the dark-sector physics, is thus to test this equivalence principle. The possibilities are:

- *Test of Lorentz invariance*, using the propagation of high-energy particles.
- *Test of local position invariance*, by testing the constancy of the non-gravitational constants (see Section 12.3 below). The detection of a variation of some constants is, in general, associated with a violation of the universality of free fall, so that tests of local physics can have a great importance in this context. Notice also that such a signature is expected in the context of models of quintessence where the quintessence field is the dilaton of string theory (see Chapter 13).
- *Test of the Poisson equation* on sub-Hubble scales [15]. If gravity is described by general relativity the matter density contrast and the gravitational potential are

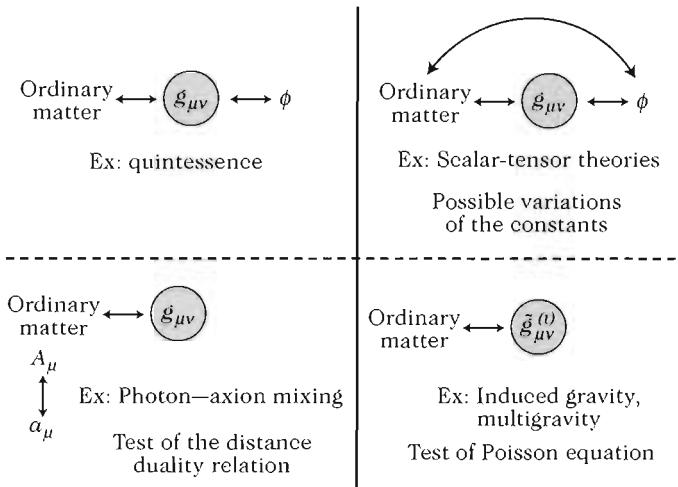


Fig. 12.1 Universality classes of the models allowing for the construction of the dark-sector physics. The classes differ depending on the nature of new fields introduced and on their coupling to matter. (top left): Models involving a new kind of matter (not necessarily, but in practice very often, a scalar field!) interacting only gravitationally (e.g., quintessence, K-essence, . . .). In this case, the only observational effect appears mainly in the equation of state of the cosmic fluid and thus on the function $H(z)$ and growth rate of density perturbations. (top right): If such a light field is non-minimally coupled to matter then it is responsible for a new long range force. If this force cannot be screened then we have a modification of general relativity on large scales (e.g., extended quintessence). These models are distinguishable by the possible variation of some constants. (bottom right): Some models propose a more radical modification of general relativity with the introduction of an infinite number of new degrees of freedom, such as, e.g., massive gravitons (e.g., multigravity, induced gravity, . . .). In this case tests of the Poisson equation can leave a signature. (bottom left): These models introduce a new field that does not dominate the matter content of the Universe but couples to photons. The number of photons is then not conserved as they can mix with this non-observed new field (e.g., axions). All the objects observed would then be less luminous because part of their photons have oscillated into this new particle: the conclusion of acceleration from supernovae data then derives from the fact that all distances have been overestimated while the expansion may not be accelerated. In this case, the distance duality relation is violated. From Ref. [14].

related via the Poisson equation (5.17). Comparing the large galaxy catalogues with gravitational lensing effects allows for the reconstruction of both sides of this equation and thus to test it observationally. This test has been extended in many works trying to design a post- Λ CDM formalism that was roughed out in Ref. [13].

- *Test of the distance duality relation:* the relation (3.75) allows us to constrain models in which the number of photons is not conserved [14], as in the case of a photon–axion mixture.
- *Measure of the equation of state as a function of the redshift.* This measure is

unavoidable. In particular, the first question to which observations should attempt to bring an unambiguous response is: is $w \neq -1$?

- *Study of the growth rate of structure.* In the linear regime, only $H(z)$ plays a role but the study of the non-linear regime allows us to constrain deviations from general relativity [16].

12.2.2.3 Questions to be answered by a dark-sector model

The observed acceleration of the Universe has led to the problem of the nature of the dark energy and that of the cosmological constant being distinguished. This last problem is often assumed to be solved exactly ($\rho_V = 0$), assuming some yet unknown symmetry reason. The source of the acceleration of the Universe thus needs to be explained.

A physical model of dark energy must address various questions. To start with, it must explain the *coincidence problem*, i.e. why dark matter and dark energy are in comparable ratios today. It must then be capable of reproducing the equation of state of the cosmic fluid. Finally, and probably most importantly, it must provide an explanation for the origin of these fields within the framework of particle physics or in one of its extensions. Without this, a purely phenomenological approach, however useful, remains vain and purely descriptive.

12.2.3 Quintessence

We will call *quintessence* any model of dark energy involving a single scalar field, denoted by Q , with a canonical kinetic term. Clearly, if this field explains the acceleration of the Universe, it must be slow-rolling today. From (1.106), we deduce that its equation of state is

$$w(Q) = \frac{\dot{Q}^2 - 2V(Q)}{\dot{Q}^2 + 2V(Q)}, \quad (12.24)$$

that can vary between -1 and 1 . The evolution of this field is dictated by the Klein–Gordon equation

$$\ddot{Q} + 3H\dot{Q} + \frac{dV}{dQ} = 0, \quad (12.25)$$

and Q must relax towards the minimum of V .

In a quintessence model, Q has not yet reached its minimum today (otherwise we would be back to the previous problem of the cosmological constant). It must therefore be in slow-roll so that $w(Q) < -1/3$. Less diluted than matter and radiation, it would thus inevitably dominate the matter content of the Universe, although this does not explain why it starts dominating today! Moreover, the Q -domination era should have begun only recently since otherwise the model would be incompatible with observations. If the potential is steep enough for the field to roll fast in its initial phase, then it behaves as a fluid with equation of state $w = +1$, its energy is then diluted as a^{-6} , making it rapidly subdominant. Two classes of models have been constructed to take into account these constraints: (1) the ones for which the field evolves monotonously in a decreasing potential towards 0 at infinity and (2) the ones for which a scalar field, often called a pseudo-Goldstone boson, relaxes towards

its vacuum. It remains to determine which potentials permit implementation of this scenario.

12.2.3.1 Scaling solutions

Let us start by looking for the kind of potential that allows us to get a ‘constant’ equation of state, $w(Q) = n/3 - 1$. This implies that

$$\dot{Q}^2 = \frac{n}{3} \rho_Q, \quad \text{and} \quad V(Q) = \left(1 - \frac{n}{6}\right) \rho_Q. \quad (12.26)$$

In a flat Universe, during a phase dominated by a fluid with an equation of state $w = m/3 - 1$ ($m = 3, 4$, respectively, for matter and radiation) and the scalar field (see Refs. [17, 18]), the Friedmann equation, $H^2 = \kappa(\rho_0 x^{-m} + \rho_{Q0} x^{-n})/3$, can be used to rewrite the first equation in the form

$$\frac{dQ}{dx} = \frac{A}{x\sqrt{1 + B^2 x^{n-m}}}, \quad (12.27)$$

with $B^2 = \rho_0/\rho_{Q0}$, $\kappa A^2 = n$ and $x = 1/(1+z)$.

We should then distinguish between two cases. If $n = m$ then $Q - Q_0 \propto \ln x$, so that we obtain the potential

$$V(Q) \propto \exp\left(-\lambda \frac{Q}{M_P}\right), \quad (12.28)$$

where we have chosen a positive λ for the potential to satisfy $V \rightarrow 0$ when $Q \rightarrow \infty$ and have introduced the Planck mass by hand so that λ is dimensionless. A subdominant scalar field evolving in a potential (12.28) behaves as the matter that dominates the Universe. If $n \neq m$, then the solution of (12.27) now gives $Q - Q_0 \propto \ln [\sqrt{1 + B^{-2} x^{m-n}} + B x^{(m-n)/2}] / (m-n)$, so that the potential is

$$V(Q) \propto Q^{-2n/(m-n)}. \quad (12.29)$$

If $m > n$, then $V \rightarrow 0$ when $Q \rightarrow \infty$ and $w(Q) < w$ so that the scalar field will finally dominate.

To summarize, we obtain that for the two families with potentials,

$$V_1(Q) = M^4 e^{-\lambda Q/M_P}, \quad V_2(Q) = M^{4+\alpha} Q^{-\alpha}, \quad (12.30)$$

where $\lambda > 0$ and $\alpha > 0$, the quintessence field, Q , behaves as a fluid with equation of state

$$w_1(Q) = w, \quad w_2(Q) = \frac{w\alpha - 2}{\alpha + 2}, \quad (12.31)$$

respectively, as long as the Universe is dominated by a component of constant equation of state w . These solutions are also known as tracking solutions since the equation of state of the scalar field is determined by that of the fluid that dominates the matter content of the Universe. With the potential V_1 , the scalar field always corresponds to a constant fraction of the matter fluid, while with the potential V_2 it will always end by dominating.

12.2.3.2 Attractor mechanism

It is interesting to show that these so-called scaling solutions are attractors for the scalar field cosmological evolution [17,18]. To illustrate this, let us consider the example of an inverse power-law potential V_2 and set $t = \exp \tau$, $u = Q/Q_{\text{scaling}}$. The Klein-Gordon equation then takes the form

$$u'' + \left(\frac{6}{m} + \frac{4}{\alpha+2} - 1 \right) u' + \frac{2}{\alpha+2} \left(\frac{6}{m} - \frac{\alpha}{\alpha+2} - 1 \right) u(1-u^{-\alpha}) = 0. \quad (12.32)$$

Setting $v = u'$ and linearizing around the scaling solution ($u = 1 + \epsilon$), we obtain

$$\begin{pmatrix} \epsilon \\ v \end{pmatrix}' = \begin{bmatrix} 0 & 1 \\ -2 \left(\frac{6}{m} - \frac{\alpha}{\alpha+2} \right) & \left(1 - \frac{6}{m} - \frac{4}{\alpha+2} \right) \end{bmatrix} \begin{pmatrix} \epsilon \\ v \end{pmatrix}. \quad (12.33)$$

By computing the eigenvalues of this system, we obtain that the solution is an attractor if $1 - 6/m - 4/(\alpha+2) < 0$ (see Fig. 12.2).

So during the matter and radiation eras, the field Q behaves as a fluid with an equation of state depending on that of the field that dominates the matter content of the Universe. The attractor mechanism ensures that one can start from a large domain of initial conditions to reach this scaling solution. We can thus think that for redshifts lower than that of nucleosynthesis, the scaling solution has been reached. Figure 12.3 describes the evolution of the quintessence field in two different models. As expected, the field Q first behaves as a^{-6} and soon becomes subdominant before reaching the scaling regime, hence explaining the changes in its equation of state. It can finally dominate in a more recent past and explain the acceleration of the Universe.

At this level of analysis, the form of the potentials (12.30) is *ad hoc* and only justified by its cosmological interest. It remains to understand how such potentials can appear in the framework of models from high-energy physics.

12.2.3.3 Model building

Potentials decreasing monotonically towards 0 at infinity often appear in models where supersymmetry is broken dynamically. Actually, supersymmetric theories are characterized by potentials with many flat directions (directions along which the potential varies slowly). These degeneracies are lifted during the dynamical breaking of supersymmetry, which is controlled by a scalar field.

Many potentials have been proposed and we mention here a few examples:

- *Exponential potential.* It can appear in scenarios of spontaneous symmetry breaking by gaugino condensation in effective superstring theories. For large values of the dilaton, the condensate induces a potential $V(\phi) \propto \exp(-3\phi/b)$. A similar behaviour is obtained for the moduli [19]. We thus obtain a first class of potentials

$$V(Q) = M^4 \exp \left(-\frac{\lambda Q}{M_p} \right). \quad (12.34)$$

If the field dominates, its equation of state is

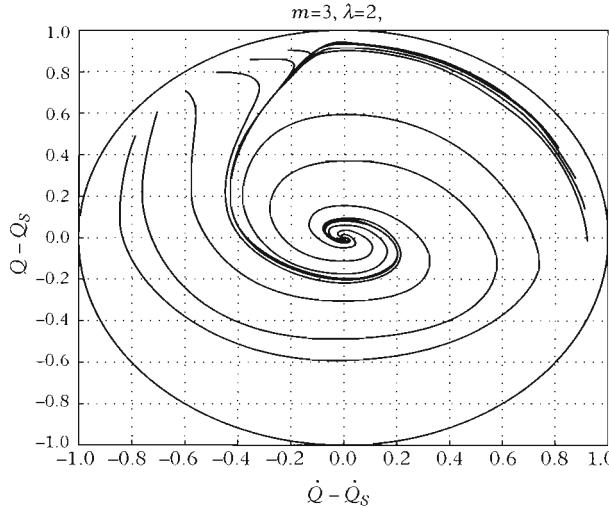


Fig. 12.2 The scaling solution is represented by the point of coordinates $(0,0)$. This phase diagram for a quintessence model with potential $V(Q) \propto \exp(-\lambda Q/M_p)$ illustrates the attractor mechanism when the matter content of the Universe is dominated by matter ($m = 3$).

$$w(Q) = \frac{\lambda^2}{3} - 1. \quad (12.35)$$

However, as we have seen, the field must initially be subdominant and its energy density then always remains a given fraction of that of the dominant fluid,

$$\Omega_Q = 3 \frac{1+w}{\lambda^2}. \quad (12.36)$$

Therefore in its minimal version, this potential does not provide a satisfying model of quintessence. Notice that the value Ω_Q is then constrained by primordial nucleosynthesis that limits the number of relativistic degrees of freedom. The constraint $\Omega_Q(1 \text{ MeV}) < 0.13$ implies that $\lambda > 5.5$.

An ‘accident’ in the potential can allow for the field to slow down and to dominate, for example, with the ‘string inspired’ potential [20],

$$V(Q) = \left[A + \left(\frac{Q}{M_p} - B \right)^\alpha \right] e^{-\lambda Q/M_p}.$$

– *Double-exponential potential.* The previous model has been extended to

$$V(Q) = M^4 \left[\exp \left(-\alpha \frac{Q}{M_p} \right) + \exp \left(-\beta \frac{Q}{M_p} \right) \right], \quad \alpha > |\beta|, \quad (12.37)$$

where the sign of β is arbitrary. As long as Q is subdominant, the first exponential dominates and $w_Q = w$. If $\beta < 0$, the potential has a minimum, $Q_{\min}/M_p =$

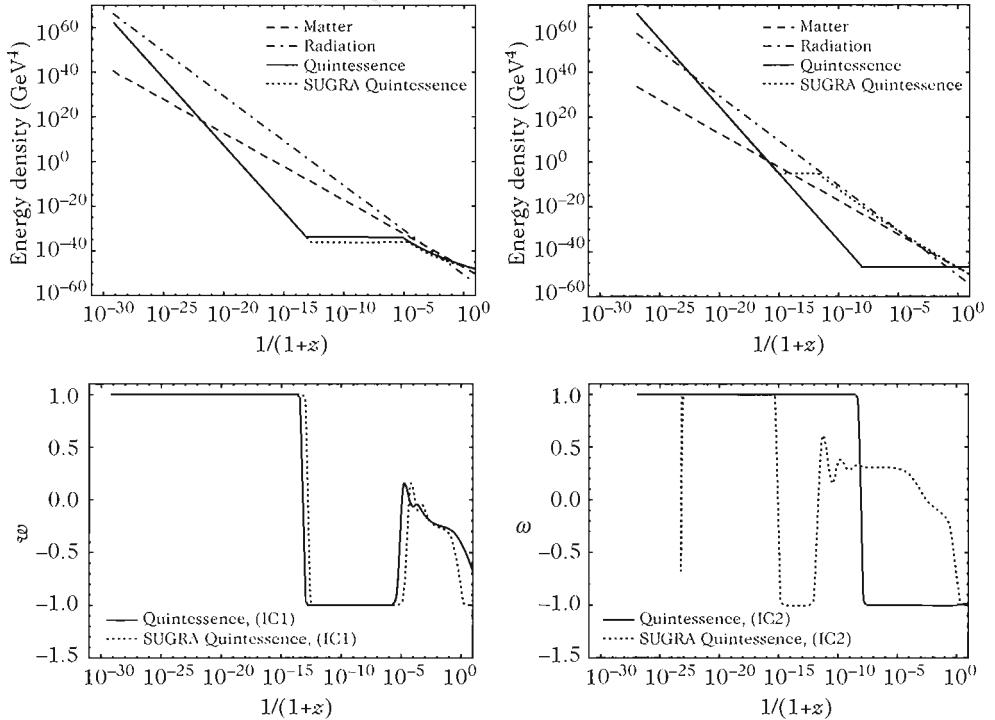


Fig. 12.3 Evolution of the energy densities of matter, radiation and quintessence field for two quintessence potentials (inverse power law with $\alpha = 6$ and SUGRA) for two sets of initial conditions (top) and corresponding equation of state (bottom). The field behaves initially as a^{-6} and becomes subdominant. Its final evolution depends on the fluid dominating the dynamics of the Universe.

$(\alpha - \beta)^{-1} \ln(-\alpha/\beta)$, towards which Q is attracted and thus $w \rightarrow -1$. Unlike a pure cosmological constant the field can oscillate before stabilizing. One should, however, explain the tuning between M , α and β that allows for $V(Q_{\min}) \sim 3M_p^2 H_0^2$. If $\beta > 0$ then one of the two exponentials eventually dominates over the other, so that, at large times, the field behaves with an equation of state given by the form (12.35).

- *Inverse power law.* This model is the archetype of quintessence models and also the oldest considered [17, 18]. As shown in the previous analysis, the field is attracted towards a scaling solution. Figure 12.4 represents the variation of its equation of state with redshift for different values of α . Unlike the exponential potential, the scalar field always dominates at large times.

A model of supersymmetry breaking in the framework of chromodynamics can provide such a potential [19]. In a QCD model, with gauge group $SU(N_c)$ with $N_f < N_c$ quark flavours, the gauge coupling becomes strong towards the supersymmetry-breaking scale and meson binding states can form. The dynamics is

then described by a potential of the form

$$W = (N_c - N_f) \frac{M^{(3N_c - N_f)/(N_c - N_f)}}{(\det \Pi)^{1/(N_c - N_f)}}.$$

Interpreting the meson field, Π , as the quintessence field, Q , its potential is then an inverse power law with exponent

$$\alpha = 2 \frac{(N_c + N_f)}{(N_c - N_f)}. \quad (12.38)$$

The dynamics of these models is summarized by Figs. 12.3 and 12.4.

- *SUGRA potential.* In the previous model, Q has a value of the order of M_p today. Corrections from supergravity [21] can no longer be neglected and transform the potential into

$$V(Q) = M^{4+\alpha} Q^{-\alpha} \exp\left(\frac{Q^2}{M_p^2}\right). \quad (12.39)$$

The effect of this correction is to stabilize Q towards M_p . The equation of state of quintessence is thus closer to -1 than for a potential with inverse power law (see Fig. 12.3).

- *Pseudo-Goldstone boson.* These models imply a field relaxing towards the *minimum* of its potential, $V = M^4 v(Q/f)$ where f is the field VEV. If the field is close to its VEV and explains the acceleration of the Universe, one needs

$$M^4 \sim H_0^2 M_p^2 \quad \text{and} \quad \frac{V''}{2} \sim \frac{M^4}{f^2} \leq H_0^2,$$

the second inequality being simply the slow-roll condition evaluated today. We infer that we need

$$f = \mathcal{O}(M_p) \quad \text{and} \quad M \sim 10^{-3} \text{ eV}, \quad (12.40)$$

so that the field must be very light, $m_Q \sim 10^{-33}$ eV, which is only natural if the field is a pseudo-Goldstone boson. A typical example is that of the axion for which

$$V(Q) = M^4 \left[1 + \cos\left(\frac{Q}{f}\right) \right]. \quad (12.41)$$

- *Extended quintessence.* The existence of the scaling solution and the attractor mechanism has been extended to the framework of tensor-scalar theories [22]. The quintessence field is then the dilaton and there can be a double-attractor mechanism, towards the theory of relativity and towards the scaling solution [23]. A model inspired from string theory, the so-called ‘runaway dilaton’, has been constructed [24, 25]. In this model, the scalar field is not universally coupled to matter, leading to a variation of the non-gravitational constants. An interesting phenomenon also appears if we assume that the quintessence field is coupled differently to ordinary matter and to dark matter [26].

12.2.3.4 Quintessential problems

The existence of the attractor mechanism towards the scaling solution relaxes the constraints on the initial conditions (but does not suppress it, as we have seen in a similar case in Chapter 8) of the scalar field. However, this does not mean that such models do not require fine tuning.

Indeed, for the field Q to explain the acceleration of the Universe, one needs $\Omega_{Q0} \sim 0.7$, which imposes a constraint on the values of the parameters of the potential. For an inverse power-law potential, we obtain the condition

$$M^{4+\alpha} \sim H_0^2 M_p^{\alpha+2}, \quad (12.42)$$

which amounts to imposing that today $V(M_p) \sim \rho_{\text{crit}}$. Figure 12.4 gives the relation between M and α , obtained numerically, for a flat Universe with $\Omega_{\Lambda0} = 0.7$. There is therefore still a tuning to be made. The good point of (12.42) is that the energy scale can now be fixed close to the TeV scale by choosing α properly.

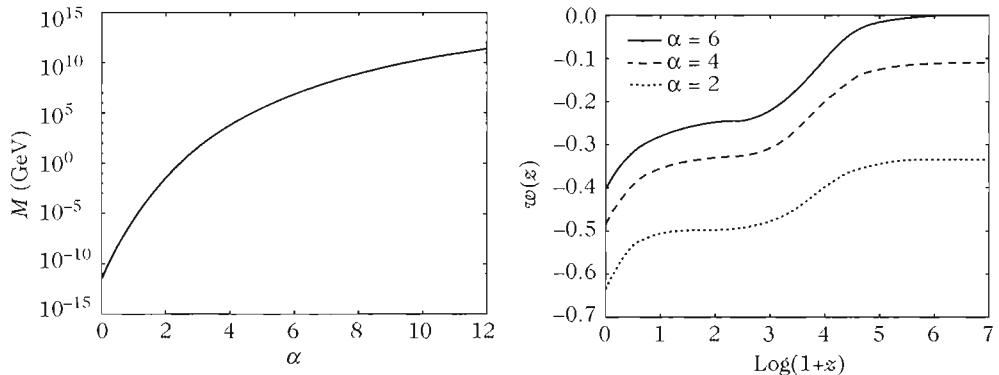


Fig. 12.4 (left): Value of the energy scale, M , required to explain the fact that dark energy represents 70% of the matter content today as a function of the exponent α for inverse power law potentials. (right): Evolution of the equation of state as a function of the redshift for different values of α .

As for the field itself, it must still be very light since the Klein–Gordon equation tells us that if $H^2 > V''$, Q is very quickly damped and if $H^2 < V''$ it is not in a slow-roll regime. Independently of the exact form of the potential, one needs

$$m_Q \sim H_0 \sim 10^{-33} \text{ eV}. \quad (12.43)$$

This mass is close to no other characteristic scale of particles physics, which makes it a naturalness problem. If this scalar field appears in the framework of a model of particle physics, then it must be coupled to other fields. It therefore generically induces a long-range force that can violate the universality of free fall. It is also difficult to imagine a symmetry that prohibits a coupling to the kinetic terms of the gauge fields, of the form $\sim (Q/M_p)^n F^2$. As we discuss later, these kinds of couplings are strongly

constrained. A solution is to assume that the quintessence field obeys a translation symmetry, $Q \rightarrow Q + \text{const.}$, but only approximatively since otherwise the potential would be absolutely flat (cf. the example of the pseudo-Goldstone boson). Another solution is to assume that the quintessence field is the dilaton and that there exists a sufficiently efficient decoupling mechanism, such as the attraction towards general relativity, (see Chapter 10) or such as the chameleon mechanism [27].

A light field must also be protected since otherwise the radiative corrections tend to induce a mass a priori larger than that observed. An interesting solution is to consider that the quintessence field is a modulus [28].

Finally, one may wonder if these models really explain the cosmological observations. Indeed, if we simply want to reproduce the observed function $E(a)$, one should be able to design a model inducing an energy density $\rho_{\text{DE}}(a)$, obtained from the observed function $H^2(a)$ by subtracting the contributions of the matter we know (i.e. pressureless matter and radiation). However, one can always construct a potential, in the parametric form, reproducing the pair $\{H(a), \rho_{\text{DE}}(a)\}$,

$$\begin{aligned} V(a) &= \frac{H(1-X)}{16\pi G_N} \left(6H + 2aH' - \frac{aHX'}{1-X} \right), \\ Q(a) &= \int \frac{d \ln a}{\sqrt{8\pi G_N}} \left[aX' - 2(1-X)a \frac{H'}{H} \right], \end{aligned} \quad (12.44)$$

with $X(a) \equiv 8\pi G_N \rho_{\text{DE}}(a)/3H^2(a)$. In some way, these models are not predictive since it is always possible to construct one that reproduces the entire equation of state $w(a)$ of dark energy!

The contribution of a model of quintessence is therefore not in its resolution of the problems presented by cosmological observations but in the fact that a model deduced from high-energy physics can also explain these observations. For now, even if there is some justification for the suggested potentials, the implementation into a high-energy framework remains difficult and a question to study in more detail.

12.2.4 Other models

12.2.4.1 Solid dark matter

In models of topological defects involving different kinds of defects or non-Abelian symmetries, there are cases where topological obstructions prevent reconnections. The defects then form a ‘solid’ web. An example is provided by an axion model with a symmetry-breaking scheme of the type $U(1) \rightarrow Z_N$. The Universe is then filled with a diffuse dark matter component with negative pressure [29].

These models give a clear prediction for the equation of state, whether we consider cosmic strings or domain walls,

$$w_{\text{wall}} = -\frac{2}{3}, \quad w_{\text{string}} = -\frac{1}{3}. \quad (12.45)$$

In the case of walls, we obtain today $\Omega \sim 1$ for a phase transition at an energy scale of the order of 100 keV. So, such defects do not induce any catastrophic signature on the cosmic microwave background. Unfortunately, the observational constraints on the equation of state rule such models out.

12.2.4.2 *K*-essence

K-essence [30, 31] considers a scalar field with non-canonical kinetic term. This model represents a transposition of the models of *K*-inflation [32] to the recent Universe. The mechanism relies on the dynamical properties of this field and on the existence of an attractor mechanism, just as in quintessence.

The action of this scalar field, minimally coupled to gravity, takes the form

$$S = \int \left[\frac{R}{16\pi G_N} - p(\phi, X) \right] \sqrt{-g} d^4x, \quad (12.46)$$

where $X \equiv \frac{1}{2}\partial_\mu\phi\partial^\mu\phi$. To simplify, one sometimes makes the decomposition $p(\phi, X) = K(\phi)\tilde{p}(X)$ with $K(\phi) > 0$ (see Ref. [33] for the general case). The scalar field then has no potential but note that if $\tilde{p}(0) \neq 0$ then, for small values of X , expanding \tilde{p} around $X = 0$ and ignoring quadratic and higher-order terms, leads to a kinetic term $\tilde{p}'(0)XK(\phi)$ and a potential $\tilde{p}(0)K(\phi)$. After a field redefinition, this leads to an ordinary scalar field with some potential.

Using the analogy with a fluid, the action (12.46) implies that

$$\rho_K = K(\phi)\tilde{p}(X), \quad P_K = K(\phi)\tilde{p}(X), \quad (12.47)$$

with $\tilde{p}(X) = 2X\tilde{p}'(X) - \tilde{p}(X)$. The speed of sound and the equation of state are then given by

$$w_K(X) = \frac{\tilde{p}(X)}{2X\tilde{p}'(X) - \tilde{p}(X)}, \quad c_s^2(X) = \frac{\tilde{p}'(X)}{\tilde{p}'(X) + 2X\tilde{p}''(X)}. \quad (12.48)$$

We can restrict ourselves to the class of models considered by imposing $\rho_K(X) > 0$, $w_K(X) \geq -1$ and/or $c_s^2(X) > 0$. The only modification of importance for the field dynamics is that of the Klein–Gordon equation that takes the form, if $\tilde{p}'(X) \neq 0$,

$$\ddot{\phi} + 3Hc_s^2(X)\dot{\phi} + \frac{K'(\phi)}{K(\phi)} \frac{\tilde{p}(X)}{\tilde{p}'(X)} = 0. \quad (12.49)$$

As an example [30], if $K(\phi) \propto \phi^{-\beta}$ and $\tilde{p}(X) = X - X^2$, then there exists a scaling (and tracking) solution that is an attractor of the dynamics with an equation of state

$$w_K(X) = -1 + \frac{(1+w)\beta}{2}, \quad (12.50)$$

during an era dominated by a fluid with equation of state w .

This example shares a strong similarity with quintessence. To better understand this link, let us consider the theory defined by the action [34]

$$S = \int \left\{ \frac{R}{16\pi G} - K(\phi) \left[\tilde{p}(\chi) + \left(\frac{1}{2}\partial_\mu\phi\partial^\mu\phi - \chi \right) \tilde{p}'(\chi) \right] \right\} \sqrt{-g} d^4x \quad (12.51)$$

which involves a new non-dynamical field χ playing the role of a simple Lagrange multiplier, in the same spirit as for the $f(R)$ theories or of the auxiliary field F in

supersymmetry (see Chapter 10 for both examples). The variation of this action with respect to the field χ leads to the equation

$$\tilde{p}''(\chi) \left(\frac{1}{2} \partial_\mu \phi \partial^\mu \phi - \chi \right) = 0,$$

so that as long as $\tilde{p}''(\chi) \neq 0$, this equation reduces to a constraint determining X as a function of χ . The theory (12.51) is thus equivalent to the initial theory (12.46). Defining

$$Q = \int^\phi \sqrt{K(\sigma)} d\sigma, \quad V(Q) = K[\phi(Q)],$$

and, assuming that \tilde{p}' is of constant sign,

$$\psi = \tilde{p}'(\chi), \quad W(\psi) = \tilde{p}(\chi) - \chi \tilde{p}'(\chi),$$

the action (12.51) takes the form

$$S = \int \left[\frac{R}{16\pi G} - \frac{1}{2} \psi \partial_\mu Q \partial^\mu Q - W(\psi) V(Q) \right] \sqrt{-g} d^4x. \quad (12.52)$$

Provided $\tilde{p} > 0$, then $w_K > -1$ implies $\psi > 0$. For the cosmological solution, the constraint equation for ψ implies that $W'(\psi) = \dot{Q}^2/2V(Q)$. In the slow-roll regime, $W(\psi)$ is almost constant so that the evolution equation of Q ,

$$\ddot{Q} + (3H\psi + \dot{\psi})\dot{Q} + V'(Q)W(\psi) = 0 \quad (12.53)$$

reduces to the classical Klein–Gordon equation. The example (12.50) is then almost equivalent to a model of quintessence with an inverse power-law potential with exponent $\alpha = 2\beta/(2 - \beta)$ for the field $\sqrt{\psi}Q$.

This construction shows that models of K -essence can reduce to models of quintessence. However, if $w_K < -1$, the construction would lead to a model of quintessence with a kinetic term with a wrong sign. In this case, a complete study of the real degrees of freedom of the theory is required to determine whether or not it is pathological.

12.2.4.3 Tachyon

Tachyon models rely on the action [35]

$$S = - \int V(T) \sqrt{-\det(g_{\mu\nu} + \partial_\mu T \partial_\nu T)} d^4x, \quad (12.54)$$

inspired from the Dirac–Born–Infeld effective action. This action can be seen as the field-theory version of the action of a one-dimensional massive relativistic particle $\mathcal{L} = -m\sqrt{1 - \dot{q}^2}$, where q is the particle position.

In the cosmological framework, the field T behaves as a fluid with pressure and density

$$\rho_T = \frac{V(T)}{\sqrt{1 - \dot{T}^2}}, \quad P_T = -V(T)\sqrt{1 - \dot{T}^2}. \quad (12.55)$$

The Klein–Gordon equation takes the form

$$\frac{\ddot{T}}{1 - \dot{T}^2} + 3H\dot{T} = -\frac{d \ln V(T)}{dT}, \quad (12.56)$$

and the Friedmann equations, for a flat Universe, become

$$H^2 = \frac{8\pi G_N}{3} \left[\frac{V(T)}{\sqrt{1 - \dot{T}^2}} + \rho \right], \quad \dot{H} = -4\pi G_N \left[w\rho + \frac{\dot{T}^2 V(T)}{\sqrt{1 - \dot{T}^2}} \right]. \quad (12.57)$$

The interest drawn by this model relies on the fact that one can decompose the density and pressure of the tachyon (12.55) as the sum of a dark-matter component

$$\rho_{T_{DM}} = \frac{V(T)\dot{T}^2}{\sqrt{1 - \dot{T}^2}}, \quad P_{T_{DM}} = 0, \quad (12.58)$$

and a component with negative pressure

$$\rho_{T_{DE}} = V(T)\sqrt{1 - \dot{T}^2}, \quad P_{T_{DE}} = -\rho_{T_{DE}}. \quad (12.59)$$

So, in the slow-roll regime, $\dot{T} \ll 1$ and the tachyon behaves as dark energy, whereas, if $V(T) \rightarrow 0$ when $T \rightarrow \infty$ and $\dot{T}^2 \rightarrow 1$ simultaneously, it can behave as dark matter. The effective equation of state today and the ratio $r = \rho_{T_{DE}}/\rho_{T_{DM}}$ are related by the relation $(1+w)(1+r) = 1$. For $r \sim 2$, this implies $w \sim -2/3$, which is marginally in agreement with observations. One of the virtues of this model is thus mainly to suggest a similar origin for two components of the dark sector, which consequently addresses the coincidence problem.

12.2.4.4 Chaplygin gas

These models [36] were originally based on a fluid with equation of state

$$P = -\frac{A}{\rho}, \quad A > 0, \quad (12.60)$$

introduced as an effective theory in hydrodynamics. Although its form is not that of a simple perfect fluid, the speed of sound remains positive and bounded

$$c_s^2 = \frac{\partial P}{\partial \rho} = \frac{A}{\rho^2}, \quad (12.61)$$

which for a negative-pressure fluid is not a trivial fact. It has been argued that, in the framework of string theory, the Nambu–Goto action for D -branes living in a $D + 2$ -dimensional space leads to such an equation of state (see Chapter 13). Note also that this equation of state is that of a tachyon (12.54) with constant potential $V = \sqrt{A}$.

Equation (12.60) has been generalized to the class of phenomenological models

$$P = -A/\rho^\alpha, \quad A > 0, \quad 0 < \alpha \leq 1, \quad (12.62)$$

and it is easy to convince ourselves that the conservation equation (3.27) implies that

$$\rho = \left[A + B a^{-3(1+\alpha)} \right]^{1/(\alpha+1)}, \quad (12.63)$$

where B is an integration constant. This matter therefore interpolates between a dark-matter kind of behaviour at large redshifts, $a^{3(1+\alpha)} \ll B/A$, and a cosmological-constant kind of behaviour. In the case $\alpha = 1$, it is interesting to note that at small redshifts, the Chaplygin gas behaves as the superposition of a cosmological constant and a fluid with equation of state $w = +1$

$$\rho \simeq \sqrt{A} + \frac{B}{\sqrt{4A}} a^{-6}, \quad P \simeq -\sqrt{A} + \frac{B}{\sqrt{4A}} a^{-6}. \quad (12.64)$$

This class of models also intends to describe dark matter and dark energy in a single framework.

12.2.5 Other approaches

12.2.5.1 Modification of general relativity

Models of extended quintessence somehow imply a modification of general relativity since they involve a long-range scalar interaction. In what follows, we want to consider models implying a deeper modification of gravity.

Various brane-inspired models have been developed (see Chapter 13 for the definition of these concepts). Let us mention the *self-tuning* mechanism in which there is a scalar field in the bulk dynamically tuning itself to reduce the value of the effective cosmological constant on the brane [37], models with several branes [38] and models of induced gravity [39] (see Chapter 13). In this last case, the Friedmann equation takes the form

$$H^2(z) = H_0^2 \left[\Omega_K (1+z)^2 + \left(\sqrt{\Omega_{r_c}} + \sqrt{\Omega_{r_c} + \Omega_m (1+z)^3} \right)^2 \right], \quad (12.65)$$

where $\Omega_{r_c} \equiv (4H_0^2 r_c^2)^{-1}$, r_c being the characteristic distance beyond which the graviton is no longer localized on the brane. It then becomes 5-dimensional, just like gravity which thus becomes weaker at large distances. The analysis of supernovae indicates that $r_c \sim 1.4H_0^{-1}$.

Multigravity models are systems for which there are two, or more, coupled metrics. One can show that the only consistent non-linear theory implying N metrics is the sum of N actions of general relativity. In the case of a model of bigravity, we assume that both ‘worlds’ simply interact via the metric coupling

$$S = \int \sqrt{-g_1} \left[\frac{R[g_1]}{2\kappa_1} + \mathcal{L}_1(g_1, \text{mat}_1) \right] + \int \sqrt{-g_2} \left[\frac{R[g_2]}{2\kappa_2} + \mathcal{L}_2(g_2, \text{mat}_2) \right] - \mu^4 \int (g_1 g_2)^{1/4} V(g_1, g_2), \quad (12.66)$$

where V is a scalar function of $g_1^{-1}g_2$. In each sector, the matter is minimally coupled to the metric (matter of type 1 with g_1) and there are no interactions between the matter of the two sectors. This model thus introduces three parameters (κ_1 , κ_2 and μ). When $\mu \rightarrow 0$, the two worlds do not interact. The field equations then take the form

$$G_{\mu\nu}^{(i)} = \kappa_{(i)} \left(T_{\mu\nu}^{(i)} + t_{\mu\nu}^{(i)} \right), \quad (12.67)$$

where $T_{\mu\nu}^{(i)}$ is the energy-momentum tensor of matter from world i ($i = 1, 2$) and $t_{\mu\nu}^{(i)}$ that associated to the coupling between both metrics, for instance,

$$t_{(1)}^{\mu\nu} = \frac{2}{\sqrt{g_{(1)}}} \frac{\delta S_{\text{int}}}{\delta g_{\mu\nu}^{(1)}} = -2\mu^4 \left(\frac{g_2}{g_1} \right)^{1/4} \left[\frac{1}{4} g_{(1)}^{\mu\nu} V(g_1, g_2) + \frac{\partial V(g_1, g_2)}{\partial g_{\mu\nu}^{(1)}} \right]. \quad (12.68)$$

The dynamical study of this theory shows that it generically leads to an acceleration phase [40]. The latter is anisotropic, which distinguishes it from other existing models.

12.2.5.2 Photon–axion oscillation

Any pseudo-scalar field, like the axion of QCD (see Chapter 7), can couple to two photons. In the presence of a magnetic field, photons and axions can oscillate between each other since they are not the eigenstates of the mass matrix.

It has been suggested that this mechanism is efficient enough to explain supernovæ observations. The latter will seem dimmer the further away they are, simply because part of their luminous energy has been converted into axions, that we do not detect.

This new approach involves a new matter field but does not assume that it dominates the matter content of the Universe and does not modify the theory of gravity. The dynamical equations of the Universe are hence not modified and the Universe would not be in acceleration despite the observational indications. Nevertheless, this model suffers from various problems. (1) The magnetic fields and the intergalactic plasma are not homogeneous, implying a loss in the oscillation coherence, translating into an achromaticity of the supernovæ light curves. This is strongly constrained observationally. (2) Evidence for the acceleration of the Universe no longer relies on the sole observation of supernovæ and it is difficult to explain how this mechanism would affect the cosmic microwave background and the gravitational lensing effects, which when combined, also point towards the existence of a cosmic acceleration. (3) This model induces a violation of the distance duality relation (see Chapter 3) that seems now to be verified observationally.

12.2.5.3 Anthropic considerations

Many versions of the anthropic principle [41] exist, some of them so weak that they are only tautologies and others so strong that they lead to absurdities or to finalistic arguments.

This principle mainly relies on the idea that some parameters strongly influence other physical phenomena (such as our existence as physical observers) and that if these parameters had another value, these phenomena would not occur or be observable. According to this approach, it is no longer a question of deriving the values of

these parameters but to show what are the consequences of their modification for the existence or for the characteristics of some physical, chemical or biological phenomena. This approach therefore does not have the same epistemological status since it is not an explanation of the values of the parameters and it amounts to showing that these values are necessary conditions for the existence of such or such phenomena.

Notice that this argument does not prove the necessity of the existence of the phenomena in question. In simpler terms, if a physical phenomena P exists and if this implies that the condition C must be satisfied, then in turn this implies that if C is not valid, then P is impossible. But nothing tells us that we should necessarily have P . This approach is based on the necessary conditions and should not be confused with an approach based on sufficient conditions. If the phenomena P is life, then this principle is called the *anthropic principle* and amounts to taking into account that the existence of an observer is an observation that should not be neglected.

The most direct constraint concerning the cosmological constant relies on the realization that if the vacuum energy were too important, then galaxies would never form. The condition $\Omega_\Lambda(z_{\text{gal}}) \leq \Omega_m(z_{\text{gal}})$ translates into

$$\frac{\Omega_{\Lambda 0}}{\Omega_{m0}} \leq (1 + z_{\text{gal}})^3 \sim 125, \quad (12.69)$$

taking $z_{\text{gal}} \sim 4$. So the cosmological constant could be larger than that observed and yet allow for the formation of galaxies. Note that this argument assumes that the amplitude of the initial perturbations has been kept constant.

This sets a constraint on the possible value of the cosmological constant but, one can ask another question such as to determine what is the most probable value that a *typical* observer would measure. A Universe with $\Omega_{\Lambda 0}/\Omega_{m0} \sim 1$ has more galaxies than a Universe where $\Omega_{\Lambda 0}/\Omega_{m0} \sim 100$. It is thus conceivable that most observers would measure a value closer to 1 than to 100. The probability to measure a vacuum energy density of value ρ_Λ can be decomposed as

$$d\mathcal{P}(\rho_\Lambda) = \mathcal{P}_*(\rho_\Lambda)\nu(\rho_\Lambda)d\rho_\Lambda, \quad (12.70)$$

where $\mathcal{P}_*d\rho_\Lambda$ is the a priori probability to have ρ_Λ and $\nu(\rho_\Lambda)$ is the mean number of galaxies forming in a given cosmological model.

All the difficulty resides in the determination of $\mathcal{P}_*(\rho_\Lambda)d\rho_\Lambda$. Various choices, including that of a constant distribution function, have been proposed. For the approach to be satisfying, one should be capable of constructing a model in which the vacuum energy can vary spatially on scales larger than the size of the observable Universe with an increment smaller than 10^{-47} GeV^4 , without introducing a new parameter of the order of 10^{-60} . A central mechanism of this construction is the model of eternal inflation that allows for the generation of spatial distribution for some physical parameters.

Note that the conclusions of this reasoning can depend on the number of parameters that are left to vary. As for the nature of dark energy, it predicts that it is a pure cosmological constant, thus, concluding with no ambiguity that

$$w_X = -1. \quad (12.71)$$

12.2.5.4 Smoothing and non-linearity of Einstein's equations

One idea, maybe the most conservative of all, concerning dark energy is that the cosmological constant appears in the Friedmann equations simply because we have not computed correctly the value of the matter density dictating, that is smoothed on large scales, the evolution of the Universe [42].

Actually, the energy-momentum tensor of the matter distribution of the Universe, $T_{\mu\nu}(t, \mathbf{x})$, has a very complex form. In particular, it is inhomogeneous and anisotropic. The solution of the Einstein equations with this energy-momentum tensor as a source would then give a space-time whose metric, $g_{\mu\nu}$, a solution of

$$G_{\mu\nu}[g] = \kappa T_{\mu\nu}(t, \mathbf{x}), \quad (12.72)$$

is very complex. The metric describing the Universe on cosmological scales is, in principle, obtained by averaging the actual metric $g_{\mu\nu}$ on large scales, so that the Friedmann–Lemaître metric is the average $\langle g_{\mu\nu} \rangle$.

In the standard approach of cosmology, this is not what is actually done. Based on the cosmological principle, we infer the most general form the energy-momentum tensor should have on large scales, $\langle T_{\mu\nu} \rangle$ (with in particular the form of a perfect fluid imposed by the symmetries), and then we solve the Einstein equations to obtain a large-scale metric $\bar{g}_{\mu\nu}$, a solution of

$$G_{\mu\nu}[\bar{g}] = \kappa \langle T_{\mu\nu} \rangle. \quad (12.73)$$

Since the Einstein equations are not linear, clearly $G_{\mu\nu}[\bar{g}] \neq \langle G_{\mu\nu}[g] \rangle$, so that the equation for the smooth metric $\bar{g}_{\mu\nu}$ actually takes the form

$$G_{\mu\nu}[\bar{g}] = \kappa \langle T_{\mu\nu} \rangle + (G_{\mu\nu}[\bar{g}] - \langle G_{\mu\nu}[g] \rangle), \quad (12.74)$$

instead of (12.73). It is therefore unlikely that $\bar{g} = \langle g \rangle$, although both should have a Friedmann–Lemaître form. The big issue in overcoming this observation is to correctly define the averages we have mentioned, to find a process to coherently average the matter and the metric and to define the physically relevant smoothing scale.

However, if the term $[G_{\mu\nu}[\bar{g}] - \langle G_{\mu\nu}[g] \rangle]$ behaves as a cosmological constant, then this would be a very satisfying explanation of the interpretation of cosmological observations. This possibility is as yet largely unexplored and we refer to Ref. [43] for recent developments.

12.2.5.5 Interpretation of data and the Copernican principle

Finally, let us recall that the cosmological observations leading to the conclusion of the recent acceleration have all been interpreted assuming a Friedmann–Lemaître space-time. These observations are localized on our past light cone so that what we interpret as a time-like history in a homogeneous and isotropic space could be interpreted as a space-like (or mixed) history in an inhomogeneous Universe. In the standard cosmological framework, the degeneracy between space and time along the past light cone is lifted by a choice of symmetry.

For example, supernovæ observations can be reproduced by a Lemaître–Tolman–Bondi Universe only filled with dust. The expansion of such a space is not accelerated.

So, the conclusion concerning the acceleration of our Universe is closely related to the cosmological principle. This reinforces the necessity to test the Copernican principle as much as possible. Recently, several possibilities have been proposed [44].

Inhomogeneous cosmological models (Chapter 3) have unfortunately been very little studied compared to their observational predictions (cosmic microwave background, BBN, etc.) and the question of whether such a model can agree with all the observations remains open. Moreover, such models necessarily imply a ‘privileged’ location for the observer.

12.2.6 Parameterization of the equation of state

The detection of a time dependence of the equation of state would be crucial information on the nature of dark energy, as it would exclude a cosmological constant. Note, however, that the knowledge of $w(a)$ does not allow us to deduce the nature of dark energy since, as we have seen, one can always construct a quintessence or tachyon model reproducing this observation. The information is thus important but not sufficient.

Let us stress that as long as we restrict ourselves to observations in the nearby Universe, the knowledge of the equation of state is sufficient as it determines the function $H(a)$ and the growth rate of perturbations, as long as general relativity is not modified and the clustering of dark energy is negligible. It is thus useful to think about the choice of this parameterization, necessary for any data analysis. In particular, constraints published in the literature are often based on this kind of parameterization.

12.2.6.1 Constant equation of state

If the equation of state of dark energy is a constant, w_X , then the transition between the deceleration and acceleration era ($\ddot{a} = 0$) happens for

$$(1 + z_a)^{3w_X} = - \frac{1}{1 + 3w_X} \frac{\Omega_{m0}}{\Omega_{X0}}. \quad (12.75)$$

As for the matter–dark energy equality, it happens for

$$(1 + z_{de})^{3w_X} = \frac{\Omega_{m0}}{\Omega_{X0}}. \quad (12.76)$$

Here again, for the acceleration to be observed, we need $1 + 3w_X < 0$. For a flat Universe with $\Omega_{\Lambda0} = 0.7$, we obtain $z_a \sim 0.73$ and $z_{de} \sim 0.37$ if $w_X = -1$. For a flat Universe, $z_a > z_{de}$ if and only if $w_X < -2/3$.

Notice that a constant equation of state $w_X < -1$ implies a cataclysmic scenario. The resolution of the Friedmann equations leads to

$$a(t) \simeq a(t_{de}) \left[(1 + w_X) \frac{t}{t_{de}} - w_X \right]^{2/3(1+w_X)},$$

where t_{de} is the cosmic time corresponding to z_{de} . So the scale factor diverges in a finite time,

$$a(t) \rightarrow +\infty, \quad t \rightarrow t_{rip} = \left(\frac{1 + w_X}{w_X} \right) t_{de},$$

as well as the Hubble parameter and the energy density $\rho(t) \propto (t - t_{rip})^{-2}$. The Universe thus reaches a curvature singularity in the future, the ‘big rip’.

If the analysis of the observations performed with the hypothesis w_X constant showed with no ambiguity that $w_X < -1$, this could simply mean that the chosen parameterization is not valid.

12.2.6.2 General parameterization

The ideal thing to do is to introduce a parameterization depending on the fewest possible parameters and reproducing the largest set of equations of state of the candidate microphysical models. A simple popular extension is a Taylor expansion,

$$w(z) = w_0 + w_1 z, \quad (12.77)$$

implying

$$\rho_X(z) = \rho_{X0}(1+z)^{3(w_0-w_1+1)} e^{3w_1 z}. \quad (12.78)$$

As illustrated in Fig. 12.5, this approximation has a very limited domain of validity. The obvious problem is that it gets worse and worse as we consider larger redshifts and it is difficult to determine its domain of validity in z .

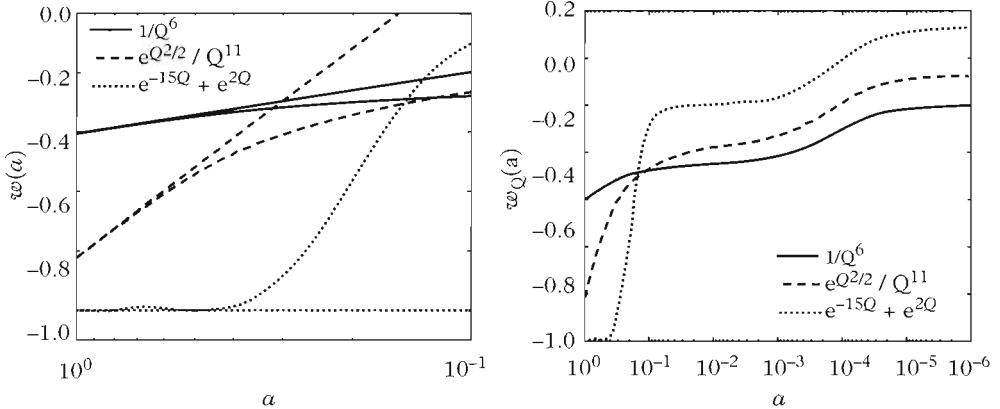


Fig. 12.5 Evolution of the equation of state for various models of quintessence as a function of the redshift. (left): The equation of state of microphysical models is compared to the linear parameterization (12.77). (right): For larger redshifts, the equation of state of dark energy has a complex structure that is well represented by a parameterization of the type (12.81).

Various improvements have been proposed, such as equations of state of the form

$$w(z) = w_0 + \frac{w_1 z}{1+z} \quad \text{or} \quad w(z) = w_0 - w_1 \log(1+z), \quad (12.79)$$

as well as a direct parameterization of the energy density

$$\rho_X(z) = \sum \frac{A_i}{(1+z)^i}. \quad (12.80)$$

The latter, introducing the exponential of the integral of the equation of state, can be more discriminating.

Numerical solutions (see Fig. 12.5) indicate that a function introducing various steps, characteristic of the different scaling solutions, can offer a good parameterization. One of them [45] introduces 7 parameters: the equations of the scaling state solution in the radiation and matter eras (w_X^r and w_X^m), the redshift at the moment where the equation of state starts changing (a_m, a_r), the duration of the transitions (Δ_m, Δ_r) and the equation of state today (w_X^0). It takes the form

$$w_X(a) = F_1 f_r(a) + F_2 f_m(a) + F_3, \quad (12.81)$$

where the functions f_i are defined by

$$f_i(a) = \left[1 + \exp\left(-\frac{a - a_i}{\Delta_i}\right) \right]^{-1}, \quad (12.82)$$

the coefficients F_1, F_2, F_3 being determined by imposing that $w_X(a)$ takes the values w_X^0, w_X^m and w_X^r today and in the matter and radiation eras.

12.2.6.3 The ‘Statefinder’ method

Another approach [46] is to introduce parameters depending on higher-order derivatives of the scale factor and generalizing the deceleration parameter. It is usual to define

$$r_0 \equiv \frac{\ddot{a}_0}{a_0 H_0^3}, \quad s_0 \equiv \frac{r_0 - 1}{3(q_0 - 1/2)}. \quad (12.83)$$

Using the Friedmann equations, we obtain

$$r = 1 + \frac{9}{2} w_X (1 + w_X) \Omega_X - \frac{3}{2} \Omega_X \frac{\dot{w}_X}{H}, \quad s = 1 + w_X - \frac{1}{3} \frac{\dot{w}_X}{w_X H}. \quad (12.84)$$

Although r and s are not independent, r depends on Ω_X, w_X and \dot{w}_X , whereas s only depends on w_X and \dot{w}_X . By increasing the number of cosmological parameters evaluated today to three (H_0, q_0, r_0), or equivalently (H_0, q_0, s_0), we can hope to better constrain the recent history of the expansion of the Universe. Notice that this method only relies on the expansion of the scale factor and therefore only assumes the validity of the description of the Universe by a Friedmann–Lemaître space-time. The Λ CDM model corresponds to $(r_0, s_0) = (1, 0)$.

12.2.6.4 Toward a post- Λ CDM formalism

As explained above, the study of the background dynamics may allow determination of $w(z)$. More information can be obtained from the large-scale structures. While a general post- Λ CDM formalism has not been designed yet, the deviations from a pure Λ CDM can be parameterized on sub-Hubble scales [13], using the notations of Chapter 5.

The Poisson equation can be generalized to

$$-k^2 \Psi = 4\pi G_N a^2 F(k, H) \rho_m \delta_m + \Delta_{de},$$

where Δ_{de} characterizes the clustering of the dark energy and F a possible long-range modification of general relativity.

The two gravitational potentials can be related by

$$\Delta(\Phi - \Psi) = \Pi_{\text{de}},$$

where Π_{de} accounts for an effective anisotropic stress. The fact that $\Phi \neq \Psi$ will leave a signature on weak lensing (see Chapter 7).

The continuity and Euler equations can be generalized to

$$\delta'_{\text{m}} + \Delta V_{\text{m}} = 0,$$

and

$$V'_{\text{m}} + \mathcal{H}V_{\text{m}} = -\Phi + S_{\text{de}},$$

where S_{de} characterizes the interaction of ordinary matter with the dark energy. In this generalized framework the Λ CDM model is the point $(F, \Pi_{\text{de}}, \Delta_{\text{de}}, S_{\text{de}}) = (1, 0, 0, 0)$.

12.2.7 Implications for the formation of the large-scale structure

Unlike a cosmological constant, a dynamical dark-energy component modifies the perturbation equations derived in Chapter 5.

In the case of a minimally coupled quintessence field (see Chapter 10 for the non-minimally coupled case), modifications of the perturbation equations amount to taking into account the energy-momentum tensor (8.124) of a scalar field in the perturbed Einstein equations (5.117)–(5.120). Defining the gauge invariant perturbation, δQ^N , as for the inflaton (8.126) this then amounts to adding an additional matter component with the characteristics

$$\begin{aligned} \delta\rho_Q^N &= \dot{Q}\delta\dot{Q}^N + \frac{dV}{dQ}\delta Q^N - \dot{Q}^2\Phi, & \delta P^N_Q &= \dot{Q}\delta\dot{Q}^N - \frac{dV}{dQ}\delta Q^N - \dot{Q}^2\Phi, \\ \rho_Q(1+w_Q)V_Q &= \dot{Q}\delta\dot{Q}^N, & \pi_{ij}^Q &= 0, \end{aligned} \quad (12.85)$$

as well as the perturbed Klein–Gordon equation

$$\ddot{\delta Q}^N + 3H\delta\dot{Q}^N + \left(\frac{k^2}{a^2} + \frac{d^2V}{dQ^2}\right)\delta Q^N = 4\dot{Q}\dot{\Phi} - 2\frac{dV}{dQ}\Phi. \quad (12.86)$$

So there is no difficulty, at least in principle, to add such a component.

12.2.7.1 Long-wavelength perturbations

In order to derive the physical consequences of these new modes, one needs to impose initial conditions on the scalar-field fluctuations on super-Hubble scales.

In most models of quintessence, the homogeneous solution was attracted towards a scaling solution, $Q_{\text{scaling}}(t)$, so that the difference between the homogeneous solution and this solution, $\delta Q(t) = Q(t) - Q_{\text{scaling}}(t)$ satisfies

$$\ddot{\delta Q} + 3H\delta\dot{Q} + \frac{d^2V}{dQ^2}\delta Q = 0. \quad (12.87)$$

If the attractor mechanism is efficient, the solution of this equation should rapidly approach 0. This equation is similar to the perturbed Klein–Gordon equation (12.86)

for super-Hubble modes ($k/aH \rightarrow 0$) when one neglects the potential term. So the solutions of the homogeneous part of (12.86) tend to 0. When taking into account the contributions from the metric perturbations, which are constant for super-Hubble modes (Chapter 5) and determined by the perturbations of the matter and radiation at early time, δQ^N is attracted towards a solution that is a linear combination of the metric perturbations.

The initial conditions of the perturbations are not important, provided the scaling solution is reached early enough. The super-Hubble modes tend towards an attractor solution, which fixes the initial conditions of the field perturbations at large scales. This attractor is, however, transitional since as soon as the term $k^2\delta Q^N/a^2$ becomes comparable with the other terms then this solution no longer exists for the perturbations.

Figure 12.6 illustrates this property for the example of a quintessence model with an inverse power-law potential.

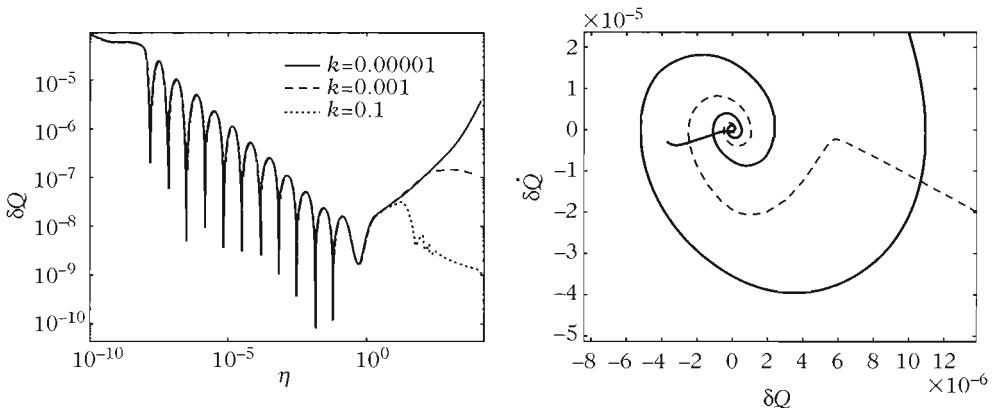


Fig. 12.6 Evolution of the perturbation of the quintessence field in a model with an inverse power-law potential. The mode has reached a scaling solution that it follows until the Laplacian term of (12.86) is no longer negligible. (left): Evolution of the different modes as a function of time. The shorter the wavelength, the earlier the mode exits the attractor solution. (right): Phase portrait representing the attraction towards a scaling regime for the perturbations for two different initial conditions. The attractor point is transitional.

12.2.7.2 Newtonian regime

Quintessence only interacts gravitationally with the other components of the Universe and is subdominant during most of the Universe history. Its effects are thus mainly felt recently, that is at low redshifts.

In a Newtonian regime, the fluctuations of the quintessence field only grow moderately since the term in $V'' > 0$ in (12.86) plays the role of a speed of sound. Thus, quintessence affects the growth of the matter density contrast mainly through the modification of the evolution of the Universe and hence of the function $H(a)$ in (5.21).

One can show that generically models of quintessence with an attractor mechanism slow down the growth of perturbations by around 20 to 30% (Fig. 12.7).

This delay in the growth has two consequences. For one, it modifies the *global* normalization of the matter power spectrum: either the spectrum is normalized with respect to the cosmic microwave background and σ_8 in a model of quintessence will be smaller than that in Λ CDM, or the normalization is performed on σ_8 in which case a model of quintessence requires larger perturbations at the time of recombination. Moreover, assuming all the quantities are normalized with respect to the cosmic microwave background, the delay in the growth of perturbations induces a delay in the time at which a given mode enters the non-linear regime. This implies a modification of the form of the non-linear spectrum. These effects can have a very significant amplitude (see Fig. 12.7).

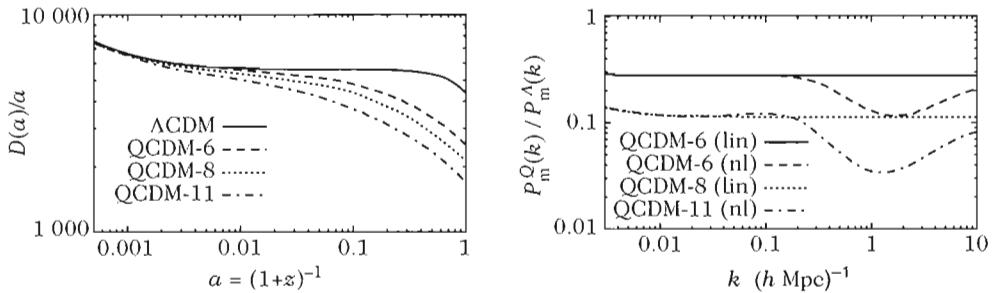


Fig. 12.7 (left): Evolution of the ratio D/a between the growing-mode solution of the density perturbations, solution of (5.21), and the growing rate in an Einstein–de Sitter space as a function of the redshift for different models. Quintessence slows down the growth of perturbations compared to Λ CDM. (right): Ratio between the matter power spectrum of a quintessence model (inverse power law) and that of Λ CDM. While the delay in the growth only affects the global normalization in the linear regime, it modifies the shape of the spectrum in the non-linear regime. From Ref. [48].

12.2.7.3 Some observational signatures

- *Cosmic microwave background*: quintessence has only a small effect on the cosmic microwave background. By modifying the cosmic fluid equation of state at small redshifts, quintessence modifies the relation between the angular distance and the redshifts, inducing a global shift in the structure of the acoustic peaks. At large angular scales, the Poisson equation tells us that in the Newtonian regime $\Delta\Phi \propto D(a)/a$. The delay in the density perturbations growth therefore manifests itself by the time-like variation of the gravitational potential at small redshifts. This modifies the amplitude of the integrated Sachs–Wolfe effect and the plateau of the cosmic microwave background anisotropies power spectrum.
- *Gravitational lensing*: gravitational lensing is one of the most powerful tools to probe the distribution as much in the linear as in the non-linear regime. Figure 12.8 illustrates how modifications of the growth of perturbations in the linear spectrum

may affect the cosmological observables. Methods of *tomography* based on the correlation of gravitational lensing effects on different source planes have, as a consequence, been suggested as promising to reconstruct the equation of state of the cosmic fluid.

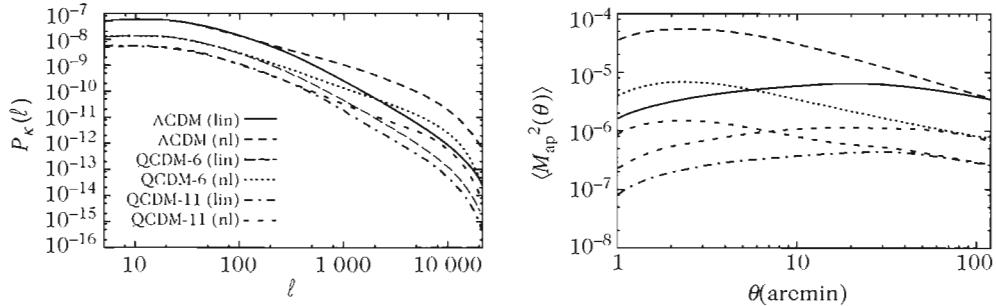


Fig. 12.8 Influence of quintessence on the gravitational lensing's observables. (left): Power spectrum of the convergence for a ΛCDM model and two models of quintessence, normalized to the cosmic microwave background. (right): Variance of the aperture mass (see Chapter 7) for the same models. Effects of the order of 80% can appear. From Ref. [48].

12.3 Varying constants

Astrophysical observations have recently relaunched a debate on the nature of the fundamental constants of physics. Detecting any variation of a fundamental constant would be, for instance, and as discussed above, a strong signal in the direction of models of quintessence and an indication for new physics.

All the references concerning this section as well as details on the observational methods can be found in Refs. [49, 50] and we refer to [51] for general considerations on the fundamental constants.

The aim of this brief section is to recall the original argument first proposed by Dirac and to show that testing the constancy of the constants of Nature is a test of the theory of general relativity. Constraints on the variation of the fine structure constant and the gravitational constant are then summarized.

12.3.0.4 Dirac argument

The question of the constancy of the constants of physics was first raised by Dirac in his ‘Large Numbers hypothesis’ that expresses the idea that very large (or very small) dimensionless numbers cannot be pure mathematical numbers. Dirac noticed that the ratio between the gravitational and electromagnetic forces between a proton and an electron, $Gm_p m_e / e^2 \sim 10^{-40}$, is of the same order of magnitude as the age of the Universe in atomic units, $e^2 H_0 / m_e c^3$. If one interprets this equality not a numerical coincidence but the reflection of a fundamental law of physics, then one can postulate that this equality is always true. This led Dirac to propose that the gravitational constant had to vary as the inverse of the cosmic time, $G \propto 1/t$.

The Dirac argument is not a theory in itself. The evolution law of the gravitational constant is postulated and no dynamical equation is proposed. The framework of tensor-scalar theories (Chapter 10) is an example of a consistent theoretical framework allowing for the formulation of this idea of a variable gravitational constant.

In a similar way, any variable constant can be modelled in a Lagrangian formalism by introducing a new dynamical variable replacing the given constant. This Lagrangian will hence provide new equations, which include the fact that this constant is dynamical, as well as an evolution equation for this ‘constant’. We stress that since any physical measurement amounts to the comparison of two physical system, only the variation of dimensionless constants can be measured. Dirac’s hypothesis actually amounts to assuming that $Gm_e m_p/\hbar c$ varies.

12.3.0.5 A fundamental test

The Einstein equivalence principle contains the hypothesis of local position invariance (Chapter 1) that implies, in its weak form, the constancy of all the non-gravitational constants. Testing the constancy of fundamental constants thus provides a test of the Einstein equivalence principle and so one of the hypotheses of general relativity, that can be performed at astrophysical scales.

Let us also stress the link with the universality of free fall. The mass of any nucleus depends on many constants, in particular on the masses of the constituents and the coupling constants, via the binding energies,

$$m(A, Z) = Zm_p + (A - Z)m_n + E_S + E_{EM}, \quad (12.88)$$

where E_S and E_{EM} are the binding energies of the strong and electromagnetic interactions. The nucleon masses can then be expressed in terms of the quark mass and the binding energies. For a macroscopic body, the gravitational binding energy should be taken into account so that the mass of any body depends on the coupling constants and on the mass ratio of the elementary particles. If one of these constants, let us call it generically α , varies, then a body moving with a speed \vec{v} with respect to the reference frame where α only depends on time undergoes an anomalous acceleration

$$\delta\vec{a} = \frac{1}{m} \frac{d(m\vec{v})}{dt} - \frac{d\vec{v}}{dt} = \frac{d \ln m}{d\alpha} \dot{\alpha} \vec{v}. \quad (12.89)$$

In an arbitrary reference frame, the value of α depends on time and space so that this acceleration takes the form

$$\delta\vec{a} = \frac{d \ln m}{d\alpha} \left(\nabla\alpha + \frac{\dot{\alpha}\vec{v}}{c^2} \right). \quad (12.90)$$

Any variation of α therefore comes with a deviation with respect to general relativity and possibly a violation of the universality of free fall since $m(\alpha)$ depends a priori on the chemical composition of the body in question. The parameter η_{12} , defined in (1.134), characterizing the violation of the universality of free fall is then given by

$$\eta_{12} = \frac{f_{ext}|f_1 - f_2|}{1 + \frac{1}{2}f_{ext}(f_1 + f_2)} \simeq f_{ext}|f_1 - f_2|, \quad (12.91)$$

for two test bodies in free fall in a external gravitational field, where the sensitivities f_i are defined by (10.22).

To put a constraint, one should know the form of the function $f_i(\phi)$, where ϕ is the field (or fields) dictating the constants' evolution, something that is only possible once the model is formulated. Considering a theory where only the fine structure constant varies, for instance, due to a coupling of the form $\frac{1}{4}B_F(\phi)F_{\mu\nu}F^{\mu\nu}$ in the Lagrangian, then $\alpha = B_F^{-1}(\phi)$. The sensitivity of any atom depends on the contribution of the electromagnetic binding energy and

$$f(A, Z) \simeq -\frac{B'_F}{B_F A m_N} \left[Z B_p + (A - Z) B_n + 98,25Z(Z-1)A^{-1/3} \text{ MeV} \right] \quad (12.92)$$

where B_n and B_p are the contributions from the electromagnetic binding energies of the neutrons and protons ($B_p = 0.63$ MeV, $B_n = -0.13$ MeV). The last term is obtained from the Bethe–Weizäcker formula for the binding energy of a nucleus. This illustrates the dependence with respect to the chemical composition.

In conclusion, any theory including dynamical constants generically induces a violation of the universality of free fall. The amplitude of the latter is related to the modifications of gravity and to the rate of variation of the constants. Given the experimental precision of the universality of free fall, it is imperative for any model of a dynamical constant to pass this test.

12.3.0.6 Constraints on the variation of the fine structure constant

The variation of the fine structure constant is constrained by many physical systems.

- *Laboratory constraints.* These methods are based on the comparison of atomic clocks using different transitions and different nuclei. Various experiments have been carried out (^{133}Cs vs ^{87}Rb , ^{133}Cs vs $^{199}\text{Hg}^+$, ^{133}Cs vs $^{171}\text{Yb}^+$, ...). For alkali nuclei, the transition of the hyperfine lines can be approximated by $\nu \propto \alpha^2(\mu/\mu_N)(m_e/m_p)Ry c F_{\text{rel}}(Z\alpha)$, where μ is the nuclear magnetic moment, μ_N the nuclear magneton, Ry the Rydberg constant, c the speed of light, and F_{rel} a function taking into account the relativistic effects that grows rapidly with Z ($d \ln F_{\text{rel}} / d \ln \alpha \simeq 0.74$ and 0.30 for ^{133}Cs and ^{87}Rb , respectively).

The comparison, during 4 years, of atomic clocks using the hyperfine transitions of the ^{133}Cs and of ^{87}Rb showed that $d \ln(\nu_{\text{Rb}}/\nu_{\text{Cs}})/dt = (0.2 \pm 7.0) \times 10^{-16} \text{ yr}^{-1}$, translating into

$$\frac{\dot{\alpha}}{\alpha} = (-0.4 \pm 16) \times 10^{-16} \text{ yr}^{-1}, \quad (12.93)$$

assuming that the magnetic moments are constant. The recent combination of several clocks (Rb, Cs, Hg+, Yb+, H) has allowed us to reach the constraints

$$\frac{\dot{\alpha}}{\alpha} = (-3.5 \pm 5.0) \times 10^{-16} \text{ yr}^{-1}. \quad (12.94)$$

- *Geochemical constraints.* The Oklo phenomena, a natural nuclear reactor that operated 2 billion years ago in Gabon, allows for the measurement of the effective cross-section of the absorption reaction $^{149}\text{Sm} + n \rightarrow ^{150}\text{Sm} + \gamma$. This absorption

has a resonance at $E_r \simeq 0.0973$ eV. Using an atomic model of the samarium nucleus, it was deduced that the resonance energy is very sensitive to small variations of α ($dE_r/d\ln\alpha \sim -1$ MeV) so that

$$\frac{\Delta\alpha}{\alpha} = (0.15 \pm 1.05) \times 10^{-7}, \quad z \sim 0.14. \quad (12.95)$$

The lifetime of unstable nuclei by α or β decay also allows us to constrain the variation of α . For instance, for the β decay, $\lambda \sim \Lambda(\Delta E)^p \propto G_F^2 \alpha^s$ where ΔE is the decay energy and Λ is a function that depends smoothly on α . For high- Z nuclei with small decay energy, $p = 2 + \sqrt{1 - \alpha^2 Z^2}$. The sensitivity to the fine structure constant is $s = p d\ln\Delta E / d\ln\alpha$. The decay of the rhenium into osmium has a sensitivity $s \sim -18\,000$. The measure of the lifetime of this nucleus in the lab and in the meteorites allows us to conclude

$$\frac{\Delta\alpha}{\alpha} = (8 \pm 16) \times 10^{-7}, \quad z \sim 0, 45. \quad (12.96)$$

- *Astrophysical constraints.* The measure of the absorption spectra by intergalactic clouds can be used to reconstruct the value of α up to redshifts of order 3. The alkali doublet method uses the fact that for fine structure doublets of alkali nuclei, $\delta\nu/\bar{\nu} \propto \alpha^2$. The analysis of 15 absorption systems of Si IV made it possible to establish that

$$\frac{\Delta\alpha}{\alpha} = (0.15 \pm 0.43) \times 10^{-5}, \quad 1.59 \leq z \leq 2.92. \quad (12.97)$$

The analysis [52] based on many multiplets had initially led to a detection, $\Delta\alpha/\alpha = (-0.54 \pm 0.12) \times 10^{-5}$ for $0.5 \leq z \leq 3$ from 128 absorption systems. However, many systematic effects had to be studied. Various analyses performed independently at the VLT have not confirmed this result [53]. Today the constraint is

$$\frac{\Delta\alpha}{\alpha} = (-0.01 \pm 0.15) \times 10^{-5}, \quad 0.4 \leq z \leq 2.3. \quad (12.98)$$

In a similar way, the analysis of the emission spectrum of O III of quasars and galaxies have established that

$$\frac{\Delta\alpha}{\alpha} = (0.7 \pm 1.4) \times 10^{-4}, \quad 0.16 \leq z \leq 0.8. \quad (12.99)$$

- *Cosmological constraints.* A variation of the fine structure constant modifies the Thompson scattering section and thus the dynamics of the recombination. Analysis from WMAP data allows us to conclude that

$$\frac{\Delta\alpha}{\alpha} = (-1.5 \pm 3.5) \times 10^{-2}, \quad z \sim 10^3. \quad (12.100)$$

Note that there are degeneracies with the cosmological parameters so that constraints stronger than 1% do not seem realistic in the near future.

The freeze-out temperature of primordial nucleosynthesis implies many constraints (Chapter 4). A variation of α mainly affects the mass difference between proton and neutron and the value of the Coulomb barriers of different nuclear reactions. Modelling is thus difficult and at the moment, the constraint is

$$\frac{\Delta\alpha}{\alpha} = (6 \pm 4) \times 10^{-4}, \quad z \sim 10^{10}, \quad (12.101)$$

but this value is to be taken with caution. The difficulty lies in the intricate structure of QCD that makes it difficult to understand the joint variation of various constants (see Ref. [55] for an example).

All these constraints are summarized in Fig. 12.9.

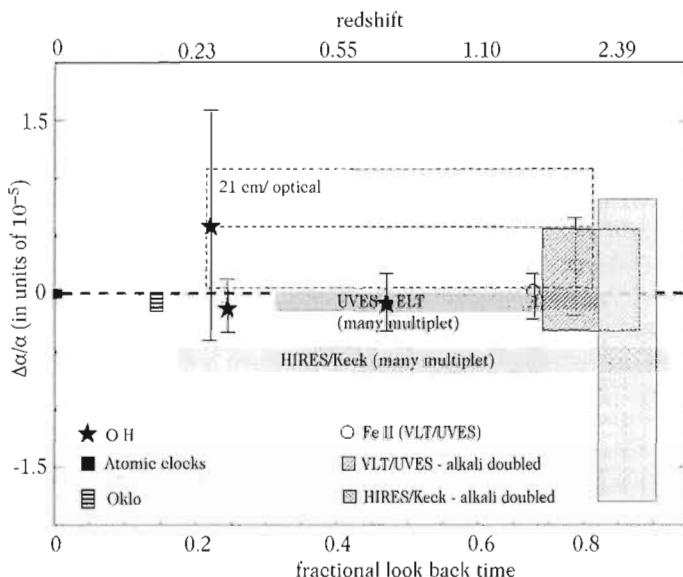


Fig. 12.9 Summary of the various constraints obtained on the variation of the fine structure constant from atomic clocks ($z = 0$), Oklo ($z \sim 0.14$), and quasar absorption spectra as a function of redshift. (Courtesy of P. Petitjean).

12.3.0.7 Other constraints

The observation of vibrorotation spectra of molecular hydrogen allows us to constrain $\mu = m_e/m_p$. The analysis of 83 systems leads to the conclusion

$$\frac{\Delta\mu}{\mu} = (2.97 \pm 0.74) \times 10^{-5}, \quad z \sim 3 \quad (12.102)$$

at 1.5σ . This value is to be taken with caution since the limiting factor is the determination of spectra in the laboratory. On the basis of recent laboratory measurements, this was improved [54] to

$$\frac{\Delta\mu}{\mu} = (2.4 \pm 0.6) \times 10^{-5}, \quad (12.103)$$

using the same astrophysical data.

The gravitational constant is constrained by primordial nucleosynthesis and in the Solar System, which implies, respectively,

$$\frac{\Delta G_N}{G_N} = 0.01_{-0.16}^{+0.20}, \quad \text{at a redshift } z \sim 10^{10}. \quad (12.104)$$

and

$$\left| \frac{\Delta G_N}{G_N} \right|_0 < 6 \times 10^{-12} \text{ yr}^{-1}, \quad (12.105)$$

today.

12.3.0.8 Theoretical models

Most models of the Universe involving extra dimensions predict dynamical constants. This is, for instance, the case of string theory (Chapter 13) where the dilaton couples directly to matter. For instance, for heterotic strings, we obtain $M_4^2 = M_H^8 V_6 \exp(-2\phi)$ and $g_{YM}^{-2} = M_H^6 V_6 \exp(-2\phi)$ for the gravitational constant and the Yang–Mills couplings, while for strings of type I we obtain $M_4^2 = M_I^8 V_6 \exp(-2\phi)$ and $g_{YM}^{-2} = \exp(-\phi)$, where V_6 is the volume of the extra dimensions. Such a coupling induces a variation of all the coupling constants.

From an effective point of view, the effect of such a dilaton can be described by an action of the type [25, 56]

$$S = \int \frac{d^4x}{16\pi G} \sqrt{-g} \left[R - 2(\partial_\mu\phi)^2 - \frac{1}{4}B(\phi)F^2 - \bar{\psi}D\psi + \dots \right]. \quad (12.106)$$

The mass of any atom depends on the function $B(\phi)$. If the function $B(\phi)$ has a maximum, then the scalar field is attracted towards this maximum, which is a minimum of the function $f_i(\phi)$, defined by (10.22). If $f_i(\phi) = \beta(\phi - \phi_m)$ then all the effects (free-fall violation, post-Newtonian parameters, constant variations) are proportional to $(\phi_0 - \phi_m)^2$. Introducing several functions $B_i(\phi)$ with maxima that do not coincide leads to effects on the violation of the universality of free fall that are, in general, larger than the experimental constraints.

12.4 Topology of the Universe

12.4.1 Local and global structures

12.4.1.1 Origin

In all the previous developments, the Universe was described by a Friedmann–Lemaître solution with locally hyperbolic, Euclidean or spherical spatial sections, depending

on the sign of the curvature. We have implicitly assumed that this was their *global* structure. All the equations we have derived, and in particular Einstein equations, are only local, hence allowing us to determine the local structure of the Universe, i.e. its *geometry*, but providing no complete information on its global structure, i.e. its *topology*. Most of our predictions were thus insensitive to this hypothesis but this does not mean we should not try to determine how we can verify this hypothesis.

References [57, 58] propose two reviews summarizing the main issues of this question.

12.4.1.2 Pro and anti

Various arguments have been presented to justify such a choice of topology, arguing that it should be trivial, in particular based on *economy* or *simplicity* principles. Furthermore, inflation predicts that the observable Universe is only a small part of the Universe (see Chapter 8). If a non-trivial topological structure existed, it would have no chance of being observable, unless there was a fine tuning in the number of e-folds. Nevertheless, it seems that the topology is capable of providing favourable initial conditions in models of inflation at low energy [59].

In turn, one can defend the existence of a topological structure basing ourselves on a *democratic* principle, according to which, assuming string theory to be valid, there should be more than four dimensions, many of which are compact (see Chapter 13). Can we then not assume that *all* the dimensions are compact and look for a dynamical mechanism to explain why three spatial dimensions are large compared to the others [60]?

There is also a preference for a space of finite volume compared to an infinite space, either based on *Mach's principle* or by putting forward conceptual problems related to this infinity. But the only simply connected manifold of finite-volume is the 3-sphere. So, if the curvature is non-positive, a finite volume manifold can only be obtained with a non-trivial topological structure.

Finally, quantum cosmology also brings arguments in favour of a finite volume Universe [61] mainly because the probability of appearance of a Universe with cosmological constant Λ and volume V goes as $\mathcal{P} \propto \exp(-\Lambda V / 8\pi\ell_p^2)$.

None of these arguments by itself can settle the question. The topological properties of space remain free and currently undetermined. We can, however, attempt to address this issue observationally. For this we need to know what are the acceptable structures (Section 12.4.2) and then determine their possible signatures on various observables (Section 12.4.4). We can then hope to either detect such a structure or to exclude its existence on scales of the size of the observable Universe.

There is another motivation to consider these Universes. The topological manifolds are not globally isotropic and have preferred directions. They offer a concrete example, in which one can compute the amplitude of these effects, and evaluate to what extent observations can signal a departure from isotropy or homogeneity of the Universe at large scales. Moreover, since most models of the primordial Universe predict a trivial topology, observing a non-trivial topology would lead to a major revolution in our theoretical perspective.

12.4.2 Mathematical introduction

In relativistic cosmology, space-time is described by a four-dimensional manifold, \mathcal{M} , (Chapter 1) that can be split as $\mathcal{M} = \mathbb{R} \times \Sigma_3$, where the spatial sections, Σ_3 , are three-dimensional spaces of constant curvature (Chapter 3). Locally, these spaces have the structure of the 3-sphere \mathbb{S}^3 if $K > 0$, of the three-dimensional Euclidean space \mathbb{E}^3 , if $K = 0$, or of the 3-hyperboloid \mathbb{H}^3 if $K < 0$. The manifolds $\mathbb{S}^3, \mathbb{E}^3$ and \mathbb{H}^3 , which we shall generically call \mathbb{X} , are called *universal covering manifolds* and are simply connected spaces with the same geometry as Σ_3 . In general, we can always express Σ_3 as the quotient of the covering space by a group Γ

$$\Sigma_3 = \mathbb{X}/\Gamma, \quad (12.107)$$

where Γ , the *holonomy group*, is a discrete subgroup of the isometry group G of the universal covering manifold, that acts freely and with no fixed point. We will denote by $|\Gamma|$ the order of this group, i.e. the number of independent elements it contains. The elements of the holonomy group are isometries, and satisfy

$$\forall x, y \in \Sigma_3, \quad \forall g \in \Gamma, \quad \text{dist}[x, y] = \text{dist}[g(x), g(y)]. \quad (12.108)$$

If the distance of a point to its image does not depend on the position of this point,

$$\forall x, y \in \Sigma_3, \quad \text{dist}[x, g(x)] = \text{dist}[y, g(y)], \quad (12.109)$$

then g is called a *Clifford translation*.

Classifying three-dimensional topological spaces with constant curvature amounts to classifying these subgroups Γ . To visualize these spaces, it is convenient to describe them with a convex *fundamental polyhedron*, \mathcal{P} , having an even number of faces. The faces of this polyhedron are associated in pairs through the generators g of the holonomy group. It can be shown that knowing the holonomy group of Σ_3 is equivalent to knowing the fundamental polyhedron, and in particular that Γ is isomorphic to $\pi_1(\Sigma_3)$. We also define the internal and external radii r_- and r_+ , respectively, as the radii of the largest and smallest sphere contained in and containing the fundamental polyhedron.

As an example, let us consider the two-dimensional torus (Fig. 12.10). It can be constructed from a square (fundamental polyhedron) with identified opposite faces. The translations bringing one face to the other are the generators of the holonomy group that is thus isomorphic to the group of the loops on the torus, $\pi_1(\mathbb{T}^2) = \mathbb{Z} \times \mathbb{Z}$ (see Chapter 10).

12.4.3 Classification of the three-dimensional manifolds

12.4.3.1 Euclidean spaces

The isometry group of \mathbb{E}^3 is $G = \mathbb{R}^3 \times \text{SO}(3)$. Any isometry can thus be decomposed as the product of a rotation/reflection, M , followed by a translation, \mathbf{T}

$$\mathbf{x} \longmapsto M\mathbf{x} + \mathbf{T}. \quad (12.110)$$

One can enumerate the 17 topological spaces by classifying the matrices M and the translations \mathbf{T} [62].

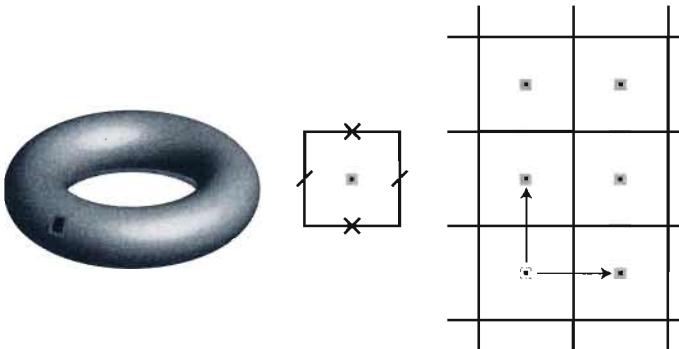


Fig. 12.10 (left): Illustration of the description of a topological manifold on the example of the two-dimensional torus, \mathbb{T}^2 . (right): This manifold can be described by a square with identified opposite faces (middle). In the universal covering space, the two translation generators of the holonomy group brings one copy of the fundamental polyhedron to a neighbouring copy.

- **Compact orientable spaces:** there are 6 compact and orientable spaces. Their fundamental polyhedrons are presented in the upper panel of Fig. 12.11. They are constructed from the torus, which is a covering space of all these spaces.
- **Torus:** the torus is the quotient of the group \mathbb{E}^3 by the group generated by three independent translations. Its fundamental domain is thus a parallelepiped. For the sake of simplicity, we only consider the *rectangular torus* generated by the three translations

$$\mathbf{T}_1 = (L_x, 0, 0), \quad \mathbf{T}_2 = (0, L_y, 0), \quad \mathbf{T}_3 = (0, 0, L_z), \quad (12.111)$$

and the *hexagonal torus* defined by

$$\begin{aligned} \mathbf{T}_1 &= (L, 0, 0), & \mathbf{T}_2 &= (-L/2, \sqrt{3}L/2, 0), \\ \mathbf{T}_3 &= (-L/2, -\sqrt{3}L/2, 0), & \mathbf{T}_4 &= (0, 0, L_z). \end{aligned} \quad (12.112)$$

Its fundamental domain is a rectangular prism with hexagonal basis. Three of the four translations (12.112) are not independent ($\mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3 = 0$) and eliminating one of these three translations one can choose another fundamental domain.

- **Half-, third- and sixth-turn spaces:** one can define three quotient spaces by replacing the translation \mathbf{T}_3 of the group (12.111) by

$$g_n : \begin{pmatrix} x \\ y \\ z \end{pmatrix} \longmapsto \begin{pmatrix} \cos 2\pi/n & -\sin 2\pi/n & 0 \\ \sin 2\pi/n & \cos 2\pi/n & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ L_z/n \end{pmatrix}, \quad (12.113)$$

which represents a corkscrew motion along the axis Oz with a rotation of angle $2\pi/n$. Since g_n composed n times with itself reduces to the translation

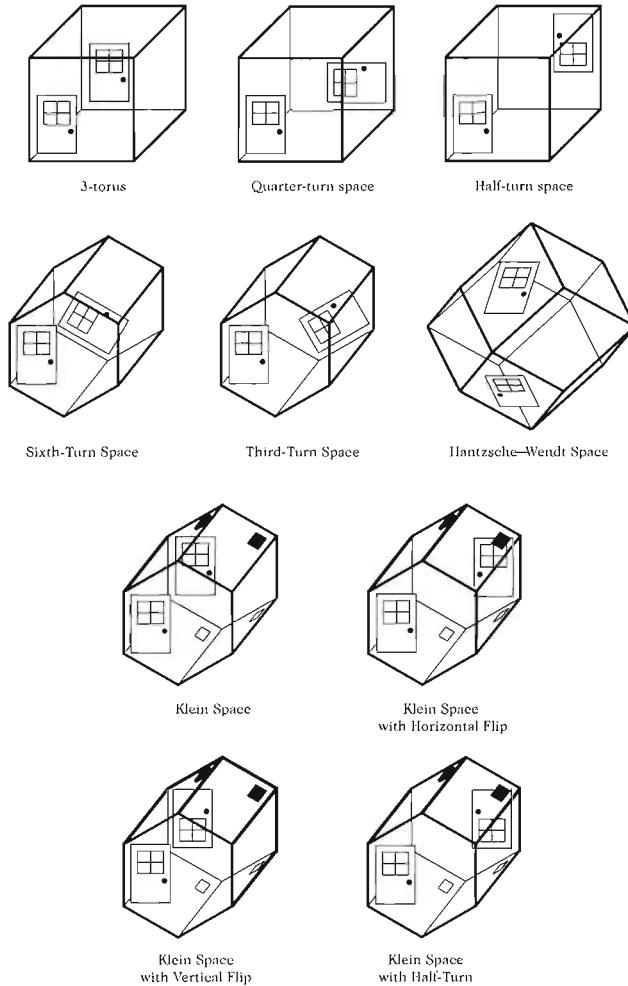


Fig. 12.11 Fundamental polyhedrons of the compact Euclidean spaces of dimension 3. There are 6 orientable spaces (top) and 4 non-orientable Klein spaces. The faces must be identified by overlapping the doors. From Ref. [62] (courtesy of Adam Weeks).

of vector \mathbf{T}_3 , n copies of these spaces form a rectangular torus. The value of n is not arbitrary and is restricted to

$$n \in \{2, 3, 6\}, \quad (12.114)$$

thereby defining the half-, third- and sixth-turn spaces, respectively.

- **Hantzsche–Wendt:** finally, another space can be constructed from a rhombic dodecahedron circumscribed in a rectangular box of size $(L_x/2, L_y/2, L_z/2)$. Its holonomy group is generated by three half-turn corkscrew motions

$$g_i : \begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} (-1)^{\gamma_i} & 0 & 0 \\ 0 & (-1)^{\alpha_i} & 0 \\ 0 & 0 & (-1)^{\beta_i} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} \alpha_i L_x/2 \\ \beta_i L_y/2 \\ \gamma_i L_z/2 \end{pmatrix} \quad (12.115)$$

with $(\alpha_i, \beta_i, \gamma_i) \in \{(1, 1, 0); (0, 1, 1); (1, 0, 1)\}$. 8 copies of this space are needed to form a rectangular torus.

- **Non-orientable compact spaces:** There are 4 non-orientable and compact spaces. Their fundamental polyhedrons are presented in the lower panel of Fig. 12.11. They are constructed from the Klein space, which is a covering space for all these spaces.

- **Klein spaces:** They are generated by two glide reflections

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} L_x/2 \\ \varepsilon L_y/2 \\ 0 \end{pmatrix}, \quad (12.116)$$

with $\varepsilon = \pm 1$ and a translation along the Oz axis,

$$\mathbf{T}_3 = (0, 0, L_z). \quad (12.117)$$

- **Quotient spaces:** One can generate three quotient spaces of the Klein space by replacing the translation (12.117) with the transformation

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} \varepsilon_1 & 0 & 0 \\ 0 & \varepsilon_2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ L_z/2 \end{pmatrix}, \quad (12.118)$$

$(\varepsilon_1, \varepsilon_2) = (-1, 1)$ and $(\varepsilon_1, \varepsilon_2) = (1, -1)$ corresponding to a Klein space with, respectively, horizontal and vertical flip, while $(\varepsilon_1, \varepsilon_2) = (-1, -1)$ corresponds to a Klein space with a half-turn. Two copies of these spaces form a Klein space (since g^2 is the translation of vector \mathbf{T}_3).

- **Chimney spaces:** these spaces have two compact dimensions and their fundamental polyhedrons are represented in the upper panel of Fig. 12.12.

- **Chimney space:** like the torus, the chimney space is the quotient of \mathbb{E}^3 with the group generated by two independent translations. So this space is not compact and its fundamental domain is an infinite chimney with parallelogram sections. For the sake of simplicity, we only consider the *rectangular chimney* generated by the two translations

$$\mathbf{T}_1 = (L_x, 0, 0), \quad \mathbf{T}_2 = (0, L_y, 0). \quad (12.119)$$

- **Chimney with half-turn:** this space is generated by replacing the translation along \mathbf{T}_2 by

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ L_y/2 \end{pmatrix}. \quad (12.120)$$

Two copies of this space reproduces a simple chimney space.

- **Chimney with flip:** the translation parallel to \mathbf{T}_2 can be replaced by

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} \varepsilon_1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \varepsilon_2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} 0 \\ L_y/2 \\ 0 \end{pmatrix}, \quad (12.121)$$

$(\varepsilon_1, \varepsilon_2) = (1, -1)$ and $(\varepsilon_1, \varepsilon_2) = (-1, 1)$ corresponding to a chimney space with, respectively, vertical and horizontal flip, while $(\varepsilon_1, \varepsilon_2) = (-1, -1)$ corresponds to a chimney space with a half-turn and a flip. Two copies of these spaces produce a chimney space (g^2 is the translation of vector \mathbf{T}_2).

- **Slab spaces:** these spaces have only one compact dimension and their fundamental polyhedrons are represented in the lower panel of Fig. 12.12.
 - **Slab space:** this space is the quotient of \mathbb{E}^3 with the group generated by one translation

$$\mathbf{T}_1 = (0, 0, L_z). \quad (12.122)$$

- **Slab space with flip:** two copies of such a space produce a slab space. It is obtained by replacing the translation along \mathbf{T}_1 by

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ L_z/2 \end{pmatrix}. \quad (12.123)$$

12.4.3.2 Spherical spaces

The 3-sphere can be embedded in a four-dimensional Euclidean space. Introducing the coordinates

$$x_0 = \cos \chi, \quad x_1 = \sin \chi \sin \theta \sin \varphi, \quad x_2 = \sin \chi \sin \theta \cos \varphi, \quad x_3 = \sin \chi \cos \theta, \quad (12.124)$$

with $0 \leq \chi \leq \pi$, $0 \leq \theta \leq \pi$ and $0 \leq \varphi \leq 2\pi$, it is described by the submanifold

$$\delta^{\mu\nu} x_\mu x_\nu = x_0^2 + x_1^2 + x_2^2 + x_3^2 = +1, \quad (12.125)$$

generalizing the definition of a sphere; note that similar embedding allows us to define n -spheres with arbitrary dimensions. The distance between two points of the 3-sphere is obtained from the scalar product $x^\mu y_\mu$ (notice that here the metric is $\delta_{\mu\nu}$),

$$\text{dist}[x, y] = \arccos(x^\mu y_\mu). \quad (12.126)$$

The volume, in units of the curvature radius, contained in a sphere of radius χ is

$$\text{Vol}(\chi) = \pi (2\chi - \sin 2\chi). \quad (12.127)$$

A convenient way to visualize \mathbb{S}^3 is to consider this space as composed of two balls of \mathbb{E}^3 whose surfaces are glued together (Fig. 12.13). Each point on the surface of one of these balls is identical to a point on the surface of the other ball. To represent a point with coordinates x_μ in the three-dimensional space, we only consider the three

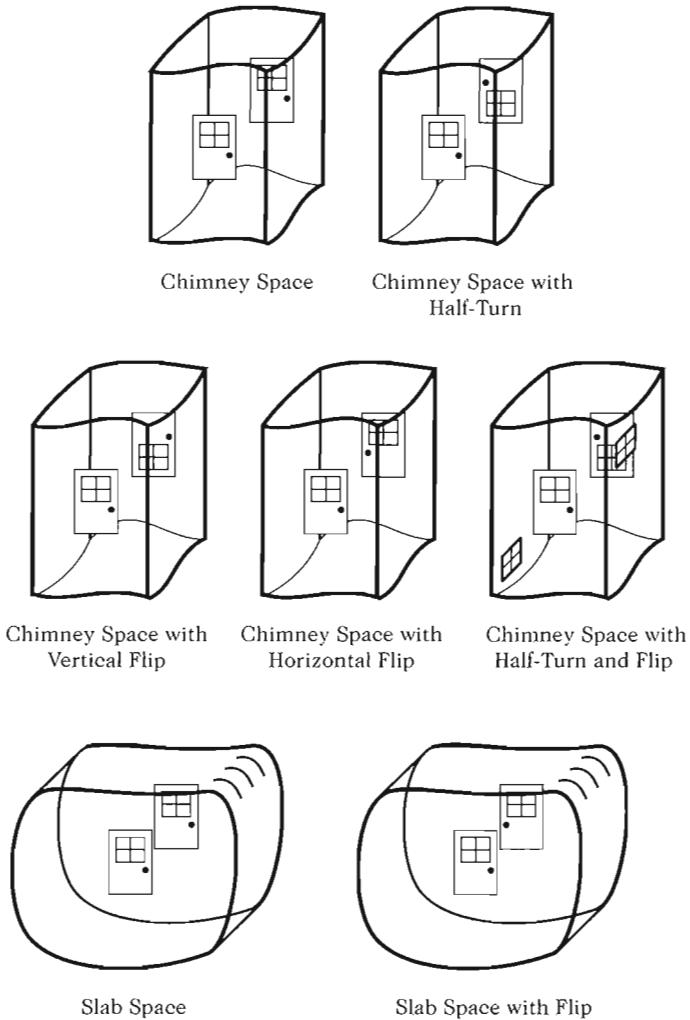


Fig. 12.12 Fundamental polyhedrons of chimney spaces (top) and slab spaces (bottom) distinguishing between orientable (left) and non-orientable (right) spaces. The faces should be identified so that the doors match. From Ref. [62] (courtesy of Adam Weeks).

coordinates $(x_i)_{i=1\dots 3}$ that correspond to a point inside the ball. However, the two points (χ, θ, φ) and $(\pi - \chi, \theta, \varphi)$ correspond to (x_0, x_1, x_2, x_3) and $(-x_0, x_1, x_2, x_3)$, explaining why two balls are required, one for $x_0 \geq 0$ and the other for $x_0 \leq 0$.

The isometry group of the 3-sphere is the four-dimensional rotation group $G = \text{SO}(4)$. Notice that for any spherical space, the volume is given by

$$\text{Vol}(\mathbb{S}^3/\Gamma) = \frac{\text{Vol}(\mathbb{S}^3)}{|\Gamma|} = \frac{2\pi^2}{|\Gamma|}, \quad (12.128)$$

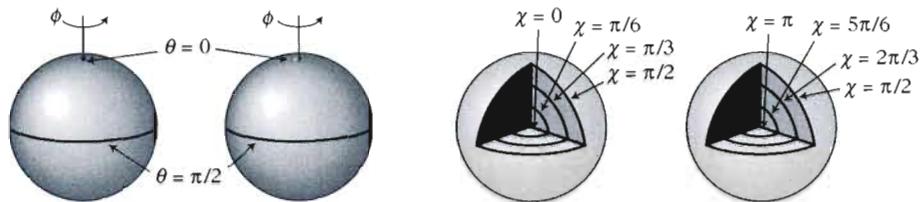


Fig. 12.13 Representation of \mathbb{S}^3 by two balls of \mathbb{R}^3 glued together. The coordinates θ and φ are the usual angular coordinates. The radial coordinate, χ , runs from 0 at the centre of one of the balls ('north pole') through $\pi/2$ at the ball's surface (spherical equator of \mathbb{S}^3) to π at the centre of the second ball ('south pole'). From Ref. [63].

and therefore, \mathbb{S}^3 is the largest of all spherical spaces: the larger the order of the group, the smaller the space.

The classification of the spherical spaces relies on the determination of the subgroups of $\text{SO}(4)$. Any isometry of $\text{SO}(4)$ with no fixed point takes the form

$$M(\theta, \phi) = \begin{pmatrix} \cos \theta & -\sin \theta & 0 & 0 \\ \sin \theta & \cos \theta & 0 & 0 \\ 0 & 0 & \cos \phi & -\sin \phi \\ 0 & 0 & \sin \phi & \cos \phi \end{pmatrix}, \quad (12.129)$$

so that all the three-dimensional spherical spaces are orientable. The isometries $M(\theta, \pm\theta)$ are Clifford translations that we describe as right ($\phi = \theta$) and left ($\phi = -\theta$). Any matrix (12.129) can be decomposed as the product of a right and a left Clifford translation as

$$M(\theta, \phi) = M(\alpha, \alpha) M(\beta, -\beta) = M(\beta, -\beta) M(\alpha, \alpha) \quad (12.130)$$

with $\alpha \equiv (\theta + \phi)/2$ and $\beta \equiv (\theta - \phi)/2$.

One can show [63] that $\text{SO}(4)$ is isomorphic to $\mathcal{S}^3 \times \mathcal{S}^3 / (\pm \{\mathbf{1}, \mathbf{1}\})$ where \mathcal{S}^3 is the quaternion group of unit length and $\mathbf{1}$ the identity quaternion (the group of the unit length quaternions is to the 3-sphere what the complex number group is to the circle). The problem now amounts to the determination of all the subgroups of \mathcal{S}^3 , and this can be simplified further by noticing that there is a homomorphism from \mathcal{S}^3 to $\text{SO}(3)$, the rotation group of the 2-sphere. The finite subgroups of $\text{SO}(3)$ are known:

- the cyclic groups, Z_n , of order n , generated by the rotations of angle $2\pi/n$ about some arbitrary axis.
- the dihedral groups, D_m , of order $2m$, generated by the rotations of angle $2\pi/m$ about some arbitrary axis together with a half-turn about a perpendicular axis.
- the tetrahedral, T , octahedral, O , and icosahedral, I , groups of respective order 12, 24 and 60 consisting of all orientation-preserving symmetries of the tetrahedron, the octahedron and the icosahedron.

The subgroups of \mathcal{S}^3 are deduced from this classification, but noticing that the homomorphism is not an isomorphism but a 2-1 relation. As a consequence, the subgroups of \mathcal{S}^3 are

- the cyclic groups, Z_n , of order n .
- the binary dihedral groups, D_m^* , of order $4m$.
- the binary tetrahedral, T^* , octahedral, O^* , and icosahedral, I^* , groups of order 24, 48 and 120, respectively.

The dihedral groups are the double covers of the spaces we consider.

From (12.130), the subgroups of $\text{SO}(4)$ are then obtained by combining the two subgroups R and L , of S^3 as

$$\Gamma = R \times L, \quad (12.131)$$

a group acting as a left (L) Clifford translation and the other as a right (R) one. This composition is actually not arbitrary and there are some impossibilities as one should make sure that Γ has no fixed point. We therefore distinguish between:

- *single-action* manifolds, obtained by considering the action of a single group of S^3 . They contain the lens (Z_n), prism (D_m^*), tetrahedral (T^*), octahedral (O^*) and dodecahedral (I^*) spaces, the latter being better known under the name of a Poincaré group. Respectively, n , $4m$, 24, 48 and 120 copies of these spaces are needed to tile S^3 . The fundamental domains of these spaces are, respectively, a lens (Fig. 12.14), a $2m$ -sided prism, a regular octahedron, a truncated cube and a regular dodecahedron.
- *double-action* manifolds are obtained by combining the action of one of the group of S^3 with the cyclic group Z_n under the condition that n and the order of this group are relative primes. This generates lens spaces of the form $L(mn, q)$ in the case of two cyclic groups (Z_n and Z_m) and $L(n, 1)$ otherwise.
- *linked action* manifolds obtained by combining the action of two groups but only allowing each element $r \in R$ to pair with a restricted subset of elements $l \in L$ to avoid fixed points.

Lens spaces, $L(p, q)$, introduced in this classification are constructed by identifying the lower face of a lens-shaped solid with its upper face after rotation of $2\pi q/p$ (Fig. 12.14). For this p and q should be relative primes and $0 < q < p$. The faces of the lens are localized on great 2-spheres of S^3 , filling half of it. One needs p copies of this space to tile S^3 . Notice that one single group can give rise to two different lens spaces. For instance, Z_5 generates $L(5, 1)$ and $L(5, 2)$.

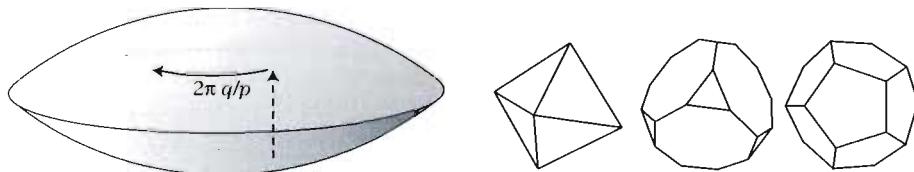


Fig. 12.14 Fundamental domains of lens spaces (left) and of the binary tetrahedral (T^*), octahedral (O^*) and icosahedral (I^*) spaces. From Ref. [63].

12.4.3.3 Hyperbolic spaces

The classification of hyperbolic spaces is much harder and is so far not completely known. The isometry group of \mathbb{H}^3 is $G = \text{SL}(2, \mathbb{C})/\mathbb{Z}_2$. Unlike what happens for spherical spaces, the volume of hyperbolic spaces increases with the complexity of the holonomy group. One can show that the volume of these manifolds must satisfy the constraint

$$\text{Vol}(\Sigma_3) \geq 0.166, \quad (12.132)$$

in units of the curvature radius. The classification of the known spaces, as well as their geometrical properties can be obtained from the freely available software SnapPea [64]. The smallest space known at the moment is the Weeks space whose volume is 0.94272.

12.4.4 Observational signatures

The choice of a topology corresponds to changing the boundary conditions of all the functions and fields evolving in this Universe.

Various attempts to detect and constrain the topology of the Universe have been explored, with varying success.

12.4.4.1 Catalogues of galaxies

If the topology has a shorter scale than that of the observable Universe, we can hope to detect multiple images of the same astrophysical object as there are several geodesics joining this object to the observer.

Initially people tried, with no success, to detect the topological images of the Milky Way and of the Coma and Abell clusters. This made it possible to establish that $r_+ \geq 15h^{-1}$ Mpc, $r_+ \geq 60h^{-1}$ Mpc and $r_+ \geq 600h^{-1}$ Mpc, respectively, assuming a Euclidean space. The main difficulty in this approach is that two topological images of the same object do not necessarily correspond to the same phase of evolution. We should therefore recognize objects whose morphology can be different. Besides the evolution, there is also no guarantee that one would observe the objects from the same point of view, so they can appear once full-front and once from the side, etc. Direct recognition methods have thus been abandoned.

Statistical methods were then developed. They rely on the property that in a catalogue containing a large number of objects, the topological images of the same object are related by holonomies that are isometries. Some distances, associated to the characteristic sizes of the fundamental polyhedron, should therefore appear more often than in the same catalogue for a Universe with trivial topology. The *crystallographic method* studies the distribution of the three-dimensional separations of all pairs of objects in a given catalogue. In practice, this method is not easy to implement. First, observations give access to an angular position and a redshift but we should determine three-dimensional distances, which requires us to know the cosmological parameters to a good precision. Secondly, the objects are not strictly comoving, but moving with a proper motion so that they are shifted with respect to the topological image's position. Finally, the method mainly emphasizes Clifford translations (as all the objects contained in the fundamental domain introduce the same distance).

All these considerations explain why these methods have not really succeeded in establishing strong constraints on the topology of the Universe.

12.4.4.2 Eigenmodes of the Laplacian

As shown in detail in Chapter 5, the equations of evolution of cosmological perturbations reduce to a set of partial differential equations introducing only a Laplacian and time derivatives. This system is easily solved in Fourier space.

12.4.4.3 Formulation of the problem

When treating a space with a non-trivial topology, one should make sure that the fields living in this space satisfy the correct boundary conditions. Any function or field, f from \mathbb{X} into \mathbb{X} , satisfying the periodicity conditions

$$\phi \circ g(\mathbf{x}) = \phi(\mathbf{x}), \quad \forall g \in \Gamma, \quad \forall \mathbf{x} \in \mathbb{X} \quad (12.133)$$

is called Γ -periodic. Such a function can be identified with a function from \mathbb{X}/Γ into \mathbb{X}/Γ .

A Γ -periodic function can be expanded in Fourier modes in a basis of eigenfunctions of the Laplacian only if these eigenfunctions are also Γ -periodic. The task is thus to determine the functions satisfying

$$\Delta \Upsilon_{k,s} = -(k^2 - K) \Delta \Upsilon_{k,s}, \quad \forall g \in \Gamma \quad \Upsilon_{k,s} \circ g = \Upsilon_{k,s}. \quad (12.134)$$

The eigenvalue $-(k^2 - K)$ depends solely on the wave number k and s represents an index (or set of indices) characterizing the solutions of the same eigenspace. The Laplacian Δ is defined from the spatial metric of curvature K .

The solutions of the Helmholtz equation are known for the universal covering spaces ($\Gamma = \{\text{Id}\}$), see Section B.5 of Appendix B from where we use the conventions. We can thus formally expand the Γ -periodic eigenfunctions on the basis of the eigenfunctions of the universal covering space as

$$\Upsilon_{k,s} = \sum_{\ell=0}^{\ell} \sum_{m=-\ell}^{\ell} \xi_{k\ell m}^s Q_{k\ell m}. \quad (12.135)$$

For each value of k , the eigenspace of the Γ -periodic eigenfunctions is a subspace of the eigenfunctions (B.70). The computation of the coefficients $\xi_{k\ell m}^s$ allows us to calculate all the cosmological observables in a Universe with non-trivial topology.

12.4.4.4 Euclidean spaces

The eigenmodes of the 17 Euclidean spaces are detailed in Ref. [62]. Here, as an example, we shall only consider the rectangular torus. In Cartesian coordinates, the eigenmodes of the Laplacian of \mathbb{E}^3 are simply plane waves, $Q_{\mathbf{k}}(\mathbf{x}) \propto \exp(i\mathbf{k} \cdot \mathbf{x})$. For a rectangular torus (12.111), the wave numbers are quantized so that they are given by

$$\mathbf{k} = 2\pi \left(\frac{n_x}{L_x}, \frac{n_y}{L_y}, \frac{n_z}{L_z} \right). \quad (12.136)$$

Using (B.21), the plane waves can be expressed in terms of the eigenmodes in spherical coordinates (B.70) as

$$Q_{\mathbf{k}}(\mathbf{x}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \left[i^{\ell} Y_{\ell m}^*(\hat{\mathbf{k}}) \right] Q_{k\ell m}(\mathbf{x}). \quad (12.137)$$

We infer that the coefficients $\xi_{k\ell m}^s$ (12.135) are given explicitly by

$$\xi_{k\ell m}^{\hat{\mathbf{k}}} = i^{\ell} Y_{\ell m}^*(\hat{\mathbf{k}}), \quad (12.138)$$

where the index s can be chosen as $\hat{\mathbf{k}}$ defined by (12.136), i.e. by the triplet (n_x, n_y, n_z) , associated with the mode (12.136).

For the other spaces, one can simply notice that the action of each isometry, g , defined by M and T , transforms a plane wave into another plane wave

$$Q_{\mathbf{k}}(\mathbf{x}) \mapsto e^{i\mathbf{k} \cdot T} e^{i\mathbf{k} \cdot M\mathbf{x}} = e^{i\mathbf{k} \cdot T} Q_{\mathbf{k}M}(\mathbf{x}). \quad (12.139)$$

If n is the smallest integer such that $\mathbf{k} = \mathbf{k}M^n$ then (i) the space generated by $\{Q_{\mathbf{k}}, \dots, Q_{\mathbf{k}M^{n-1}}\}$ is invariant under the action of g and (ii) the action of g fixes a unique invariant element of the form $a_0 Q_{\mathbf{k}} + \dots + a_{n-1} Q_{\mathbf{k}M^{n-1}}$ with $a_{j+1} = \exp(i\mathbf{k}M^j T)a_j$. Knowing (M, T) for all spaces, one can then use both these properties to determine the modes of the 17 Euclidean spaces.

12.4.4.5 Spherical spaces

The case of spherical spaces is more intricate, but the modes of all the lens and prism spaces can be determined analytically [65]. Without going into details, notice that the important point lies in the choice of a system of coordinates that respects the symmetries of the transformations of Γ , and in particular of the cyclic groups. It is useful to represent the 3-sphere in toroidal coordinates (Fig. 12.15) defined by

$$x_0 = \cos \chi \cos \theta, \quad x_1 = \cos \chi \sin \theta, \quad x_2 = \sin \chi \cos \varphi, \quad x_3 = \sin \chi \sin \varphi, \quad (12.140)$$

with $0 \leq \chi \leq \pi/2$ and $0 \leq \theta, \varphi \leq 2\pi$. For each value of χ , the coordinates (θ, φ) describe a torus, which reduces to a circle for $\chi = 0$ and $\chi = \pi/2$. The set of all these tori fills up \mathbb{S}^3 .

In a more general case, the eigenmodes are not known explicitly, but can be defined from those of the universal covering space

$$\Upsilon_{\mathbf{k}} = \frac{1}{|\Gamma|} \sum_{g \in \Gamma} Q_{\mathbf{k} \circ g}. \quad (12.141)$$

By linearity, this element is a solution of (12.134); it also satisfies the condition (12.133). The difficulty lies in extracting a basis of independent elements from these functions (see Ref. [65] for the explicit construction).

12.4.4.6 Cosmic microwave background

The cosmic microwave background was emitted at the same time everywhere in the Universe, so it contains non-local information. Since it represents the oldest observation to which we have access, it is an ideal tool to study the topology of the Universe.

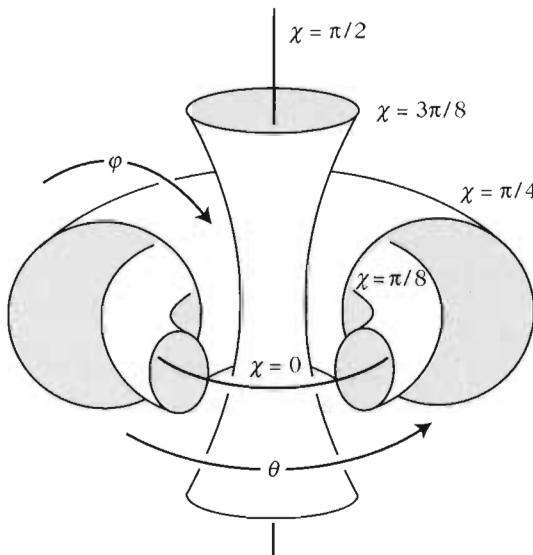


Fig. 12.15 Representation of \mathbb{S}^3 in toroidal coordinates. The 3-sphere is decomposed into slices of two-dimensional tori.

12.4.4.7 Computation of the cosmic microwave background anisotropies

As detailed in Chapter 6, the temperature fluctuations in a direction $\hat{\mathbf{n}}$ take the general form

$$\frac{\delta T}{T}(\hat{\mathbf{n}}) = \int \frac{d^3 k}{(2\pi)^3/2} O_k [Q_{\mathbf{k}}(\mathbf{x})] \sqrt{P(k)} \hat{e}_{\mathbf{k}}, \quad (12.142)$$

where $P(k)$ is the primordial spectrum, $\hat{e}_{\mathbf{k}}$ a Gaussian random variable satisfying $\langle \hat{e}_{\mathbf{k}} \hat{e}_{\mathbf{k}'}^* \rangle = \delta^{(3)}(\mathbf{k} - \mathbf{k}')$ and O_k a convolution operator depending only on k and on the cosmological parameters. Expanding $Q_{\mathbf{k}}(\mathbf{x})$ as in (B.70) and the temperature fluctuation in spherical harmonics (B.18) we obtain

$$a_{\ell m} = i^\ell \int k^2 dk \sqrt{P(k)} G_\ell(k) \hat{e}_{\ell m}(k), \quad (12.143)$$

with

$$G_\ell(k) = O_k [R_{k\ell}] \quad \text{and} \quad \hat{e}_{\ell m} = \int d^2 \hat{\mathbf{n}} Y_{\ell m}^*(\hat{\mathbf{n}}) \hat{e}_{\mathbf{k}}. \quad (12.144)$$

For a Universe with a non-trivial topology, all the relations of local physics obviously remain valid and only the boundary conditions have been changed. The only modification to make is thus to perform the substitution $Q_{\mathbf{k}}(\mathbf{x}) \rightarrow \Upsilon_{\mathbf{k}}$. Using the expansion (12.135) and noticing that the operator O_k is linear, we obtain [replacing the condition on the random variable by $\langle \hat{e}_{\mathbf{k}} \hat{e}_{\mathbf{k}'}^* \rangle = \text{Vol}(\Sigma_3) \delta_{ss'} \delta_{kk'} / (2\pi)^3$]

$$a_{\ell m} = \frac{(2\pi)^3}{\text{Vol}(\Sigma_3)} \sum_k \sqrt{P(k)} O_k [R_{k\ell}] \sum_s \xi_{k\ell m}^s \hat{e}_{\mathbf{k}}. \quad (12.145)$$

So, if the coefficients $\xi_{k\ell m}^s$ have been determined, we can produce maps of the cosmic microwave background. The $a_{\ell m}$ are Gaussian fields, since they are linear combinations of Gaussian fields, and are now correlated. Their correlation matrix is given by

$$C_{\ell m}^{\ell' m'} = \langle a_{\ell m} a_{\ell' m'}^* \rangle, \quad (12.146)$$

from which one can extract the angular spectrum

$$(2\ell + 1)C_\ell = \sum_m C_{\ell m}^{\ell m}. \quad (12.147)$$

12.4.4.8 Signatures

A non-trivial topology has three main signatures on the cosmic microwave background anisotropies.

- *Presence of correlations*: all the photons from the cosmic microwave background are emitted at the time of decoupling on the last-scattering surface. All these photons come from a 2-sphere of which we occupy the centre. In the same way a paper disk wrapped around a cylinder of smaller radius will recover itself, the surface of last scattering ‘wraps’ around the Universe and intersects itself. This intersection is a circle that appears twice on the sphere of last scattering. These two circles are physically identical so that the temperature fluctuations along these circles must be identical. This *circles in the sky* method [66] is the most direct signature of the existence of a non-trivial topology.

Figure 12.16 illustrates the existence of these correlations in the case of the torus and describes the reconstruction of the fundamental polyhedron.

This correlation would be perfect if the temperature was a scalar function on the sphere. However, this is not the case since both the Doppler effect and the integrated Sachs–Wolfe effects are noises for this method.

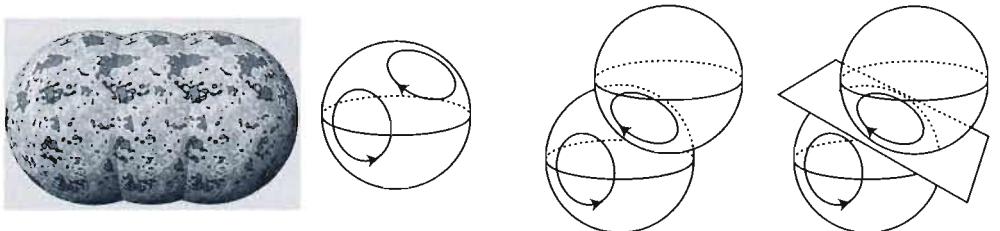


Fig. 12.16 Simulation of a cosmic microwave background map for a toroidal Universe. (left): We have represented the surface of last scattering as seen from the outside, as well as two copies by a translation of the holonomy group. These two spheres intersect on a circle along which the anisotropies are identical. (right): Detecting a pair of circles allows us to know the position of two copies of the surface of last scattering and to reconstruct the position of one of the faces of the fundamental polyhedron.

- *Violation of statistical isotropy*: the correlation matrix (12.146) is not invariant under rotations so that it does not reduce to $C_\ell \delta_{\ell\ell'} \delta_{mm'}$, as in the case of an

isotropic simply connected Universe. The presence of a correlation between different ℓ and between different m indicates a violation of the global isotropy.

Figure 12.17 represents the correlation matrix for a Universe whose topology is that of a cubic torus. Interestingly, we notice that there remains a signature of the topology even though its characteristic scale is larger than the size of the observable Universe. The difficulty lies in the development of a robust method capable of emphasising these correlations, in a significant way (so that cosmic variance would not spoil the signal). Indeed, the observations are not strictly isotropic because of mask effects (cut from the galactic plane,...) and the foreground emissions have no reasons to be isotropic.

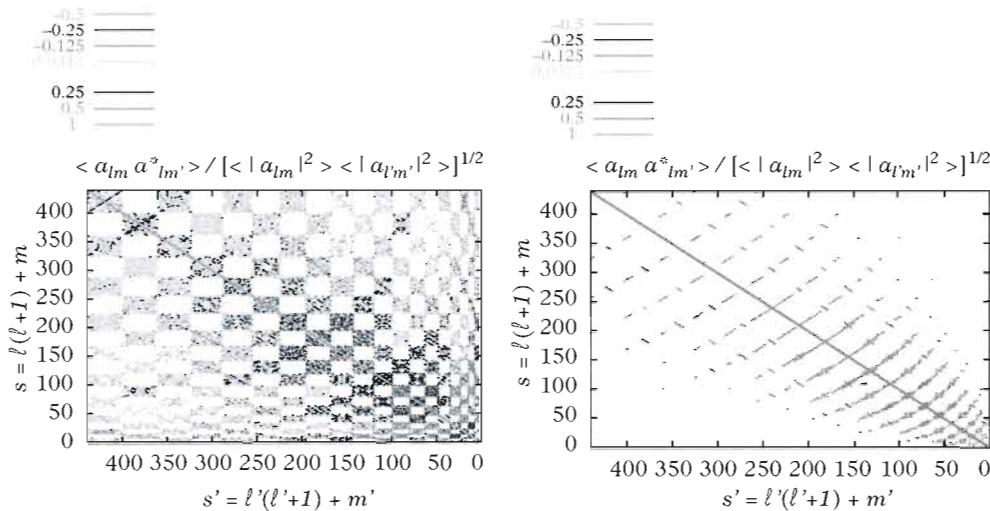


Fig. 12.17 Amplitude of the normalized correlation function $C_{\ell m}^{\ell' m'} / \sqrt{C_{\ell m}^{\ell m} C_{\ell' m'}^{\ell' m'}}$ represented as a function of $s = \ell(\ell+1) + m$ for a cubic torus. The fact that this function is not diagonal reflects the departure from global isotropy. (left): Torus of size $L = 2$ and (right): Torus of size $L = 6.5$ in units where the diameter of the last scattering surface is 6.23.

- *Modification of the angular power spectrum:* the angular power spectrum, obtained by averaging the correlation function, loses a large part of the topological information. As shown in Fig. 12.18, the topology induces a loss of power on large angular scales, mainly because the long-wavelength modes are more affected. The search for these effects is, however, not robust as it must make an assumption on the form of the primordial spectrum and they affect mainly small multipoles, for which the cosmic variance is large, so that the loss of power is usually not statistically significant. Besides, it is particularly important for very large angular scales, where the cosmic variance is large.

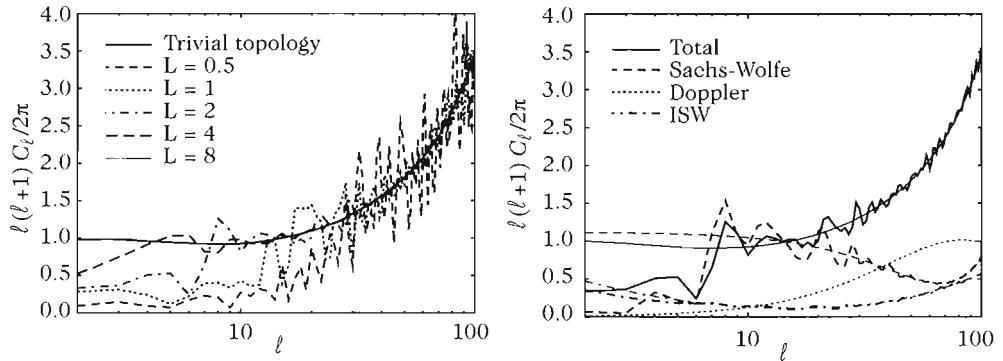


Fig. 12.18 (left): Angular power spectrum for cubic torus of different sizes (in units of Hubble radius. For these models the radius of the observable Universe is 3.17 in these units). (right): The topology mainly affects the Sachs–Wolfe term that dominates at large angular scales.

12.4.4.9 Candidates and constraints

Observations from the WMAP satellite indicate that the Universe is almost flat (Chapter 4). These observations also show that the quadrupole is smaller than expected and that there is a significant loss of power on angular scales larger than 60° . This, at the time, revived the interest in topological spaces.

Note that all the topological characteristics are expressed in units of the curvature radius. For a strictly flat Universe, there is no scale related to the curvature and it is difficult to understand why the topological scale should precisely coincide with the size of the observable Universe. For spherical spaces, on the other hand, it is the curvature scale that fixes the units and the Universe is always smaller than the 3-sphere, from which we can conclude that a spherical topology is generically observable. Notice that a positive curvature observable today is also difficult to explain in the inflationary framework.

We can also show that topological spaces explain the quadrupole observed by WMAP if they are ‘well proportioned’, i.e. if their dimensions are all of the same order of magnitude. This is the case of the Poincaré space that offers excellent agreement with observations if $\Omega_0 \sim 1.013$. This model is compelling as it adds no new parameter and predicts a curvature compatible with the constraints (it is the only parameter that controls the size of the Universe). It also predicts the existence of 6 pairs of antipodal circles of angular radius 35° that should be identified with a twist of $\pi/5$.

Searching for these circles led to no success [67]. Current analysis indicates that there are no pairs of correlated circles with an angular radius larger than 25° in the WMAP data, hence excluding the Poincaré space. Given the constraints on the cosmological parameters, this implies that the size of the Universe must be larger than 24 Gpc, whereas the size of the observable Universe is around 26.5 Gpc.

References

- [1] K. OLIVE, ‘Inflation’, *Phys. Rep.* **190**, 307, 1990.
- [2] D. LYTH and A. RIOTTO, ‘particle-physics models of inflation’, *Phys. Rep.* **314**, 1, 1999.
- [3] C. COLEMAN and E. WEINBERG, ‘Radiative corrections as the origin of spontaneous symmetry breaking’, *Phys. Rev. D* **7**, 1888, 1973.
- [4] T. BRUNIER, F. BERNARDEAU and J.-P. UZAN, ‘Radiative corrections to scalar masses in de Sitter space’, *Phys. Rev. D* **71**, 063529, 2005.
- [5] J. ROCHER and M. SAKELLARIADOU, ‘Supersymmetric grand unified theories and cosmology’, *JCAP* **0503**, 004, 2005.
- [6] L. POGOSIAN, M. WYMAN and I. WASSERMAN, ‘Observational constraints on cosmic strings: Bayesian analysis in a three dimensional parameter space’, *JCAP* **09**, 008, 2004.
- [7] R. JEANNEROT, J. ROCHER and M. SAKELLARIADOU, ‘How generic is cosmic-string formation in supersymmetric grand unified theories’, *Phys. Rev. D* **68**, 103514, 2003.
- [8] P. BINÉTRUY, G. DVALI, R. KALLOSH and A. VAN PROEYEN, ‘Fayet–Iliopoulos terms in supergravity and cosmology’, *Class. Quantum Grav.* **21**, 3137, 2004.
- [9] S. WEINBERG, ‘The cosmological constant problem’, *Rev. Mod. Phys.* **61**, 1, 1989.
- [10] P.J.E. PEEBLES and B. RATRA, ‘The cosmological constant and dark energy’, *Rev. Mod. Phys.* **75**, 559, 2003.
- [11] S. CARROLL, ‘The cosmological constant’, *Liv. Rev. Relativity* **4**, 1, 2001.
- [12] Y.B. ZEL'DOVICH, *Sov. Phys. Usp.* **11**, 381, 1988.
- [13] J.-P. UZAN, ‘The acceleration of the Universe and the physics behind it’, *Gen. Relat. Grav.* **39**, 307, 2007.
- [14] J.-P. UZAN, N. AGHANIM and Y. MELLIER, ‘The reciprocity relation from X-ray and SZ observations of clusters’, *Phys. Rev. D* **70**, 083533, 2004.
- [15] J.-P. UZAN and F. BERNARDEAU, ‘Lensing at cosmological scales: a test of higher dimensional gravity’, *Phys. Rev. D* **64**, 083004, 2001.
- [16] F. BERNARDEAU, *Constraints on higher-dimensional gravity from the cosmic shear 3-point correlation function*, [[astro-ph/0409224](#)].
- [17] B. RATRA and P.J.E. PEEBLES, ‘Cosmological consequences of a rolling homogeneous scalar field’, *Phys. Rev. D* **37**, 3406, 1988.
- [18] C. WETTERICH, ‘Cosmology and the fate of dilatation symmetry’, *Nucl. Phys. B* **302**, 668, 1988.
- [19] P. BINÉTRUY, ‘Models of dynamical supersymmetry breaking and quintessences’, *Phys. Rev. D* **60**, 063502, 1999.
- [20] A. ALBRECHT and C. SKORDIS, ‘Phenomenology of a realistic accelerating Universe using only Planck-scale physics’, *Phys. Rev. Lett.* **84**, 2076, 2000.

- [21] P. BRAX and J. MARTIN, ‘Quintessence and supergravity’, *Phys. Lett. B* **468**, 40, 1999.
- [22] J.-P. UZAN, ‘Cosmological scaling solutions of non-minimally coupled scalar fields’, *Phys. Rev. D* **59**, 123510, 1999.
- [23] N. BARTOLO and M. PIETRONI, ‘Scalar-tensor gravity and quintessence’, *Phys. Rev. D* **61**, 023518, 2000.
- [24] M. GASPERINI, F. PIAZZA and G. VENEZIANO, ‘Quintessence as a run-away dilaton’, *Phys. Rev. D* **65**, 023528, 2002.
- [25] T. DAMOUR, F. PIAZZA and G. VENEZIANO, ‘Violations of the equivalence principle in a dilaton-runaway scenario’, *Phys. Rev. D* **66**, 081601, 2002.
- [26] T. DAMOUR, G. GIBBONS and C. GUNDLACH, ‘Dark matter, time-varying G , and a dilaton field’, *Phys. Rev. Lett.* **64**, 123, 1990.
- [27] J. KHOURY and A. VELTMAN, ‘Chameleon fields: awaiting surprises for tests of gravity in space’, *Phys. Rev. Lett.* **93**, 171104, 2004.
- [28] M. PELOSO and E. POPPITZ, ‘Quintessence from shape moduli’, *Phys. Rev. D* **68**, 125009, 2003.
- [29] M. BUCHER and D. SPERGEL, ‘Is the dark matter a solid?’, *Phys. Rev. D* **60**, 043505, 1999.
- [30] T. CHIBA, T. OKABE and M. YAMAGUCHI, ‘Kinetically driven quintessence’, *Phys. Rev. D* **62**, 023511, 2000.
- [31] C. ARMENDÁRIZ-PICÓN, V. MUKHANOV and P. STEINHARDT, ‘Essentials of k-essence’, *Phys. Rev. D* **63**, 103510, 2001.
- [32] C. ARMENDÁRIZ-PICÓN, T. DAMOUR and V. MUKHANOV, ‘k-Inflation’, *Phys. Lett. B* **458**, 209, 1999.
- [33] J. GARRIGA and V. MUKHNAOV, ‘Perturbations in k-inflation’, *Phys. Lett. B* **458**, 219, 1999.
- [34] M. MALQUARTI *et al.*, ‘A new view of k-essence’, *Phys. Rev. D* **67**, 123503, 2003.
- [35] A. SEN, ‘Rolling tachyon’, *JHEP* **0204**, 048, 2002.
- [36] A. KAMENSHCHIK, U. MOSCHELLA and V. PASQUIER, ‘An alternative to quintessence’, *Phys. Lett. B* **511**, 265, 2001.
- [37] N. ARKANI-HAMED *et al.*, ‘A small cosmological constant from a large extra-dimension’, *Phys. Lett. B* **480**, 193, 2000.
- [38] R. GREGORY, A. RUBAKOV and S. SIBIRYAKOV, ‘Opening up extra-dimensions at ultralarge scales’, *Phys. Rev. Lett.* **84**, 5928, 2000.
- [39] C. DEFFAYET, G. DVALI and G. GABADADZE, ‘Accelerated universe from gravity leakage’, *Phys. Rev. D* **65**, 044023, 2002.
- [40] T. DAMOUR, I. KOGAN and A. PAPAZOGLU, ‘Non-linear bigravity and cosmic acceleration’, *Phys. Rev. D* **66**, 104025, 2002.
- [41] B. CARTER, ‘The anthropic principle and its implications for biological evolution’, *Phil. Trans. R. Soc. Lond. A* **310**, 347, 1983.
- [42] G.F.R. ELLIS, ‘Relativistic cosmology: its nature, aims and problems’, in ‘General Relativity and gravitation’, B. Bertotti *et al.* (eds.), Reidel, 1984; N. ZOTOV and W. STOEGER, ‘Averaging Einstein equations’, *Class. Quant. Grav.* **9**, 1023, 1992.
- [43] G.F.R. ELLIS and T. BUCHERT, ‘The Universe seen at different scales’, *Phys. Lett. A* **347**, 38, 2005.

- [44] J. GOODMAN, ‘Geocentrism reexamined’, *Phys. Rev. D* **52**, 1821, 1995; R.R. CALDWELL and A. STEBBINS, ‘A test of the Copernican principle’, *Phys. Rev. Lett.* **100**, 191302, 2008; J.-P. UZAN, C. CLARKSON and G.F.R. ELLIS, ‘Time drift of cosmological redshift as a test of the Copernican principle’, *Phys. Rev. Lett.* **100**, 191303, 2008.
- [45] P.S. CORASANITI and E. COPELAND, ‘A model independant approach to the dark energy equation of state’, *Phys. Rev. D* **65**, 043004, 2002.
- [46] V. SAHNI *et al.*, ‘Satefinder - a new geometrical diagnostic of dark energy’, *JETP Lett.* **77**, 201, 2003.
- [47] A. RIAZUELO and J.-P. UZAN, ‘Cosmological observations in scalar-tensor quintessence’, *Phys. Rev. D* **66**, 023525, 2002.
- [48] C. SCHIMD, J.-P. UZAN and A. RIAZUELO, ‘Weak lensing in scalar-tensor theories of gravity’, *Phys. Rev. D* **71**, 083512, 2005.
- [49] J.P. UZAN, ‘The fundamental constants and their variation: observational and theoretical status’, *Rev. Mod. Phys.* **75**, 403, 2003.
- [50] J.-P. UZAN, ‘Variation of the constants in the late and early Universe’, *AIP Conf. Proceedings* **736**, 3, 2004.
- [51] J.-P. UZAN and R. LEHOUCQ, *Les constantes fondamentales*, Belin, 2005; J.-P. UZAN and B. LECLERCQ, *Our Universe and the laws of physics; Variation in fundamental constants*, Praxis, 2008.
- [52] J.K. WEBB *et al.*, ‘Further evidence for cosmological evolution of the fine structure constant’, *Phys. Rev. Lett.* **87**, 091301, 2001.
- [53] R. SRIANAND *et al.*, ‘Limits on the time variation of the electromagnetic fine-structure constant in the low-energy limit from absorption lines in the spectra of distant quasars’, *Phys. Rev. Lett.* **92**, 121302, 2004.
- [54] E. REINHOLD *et al.*, ‘Indication of a cosmological variation of the proton-electron mass ratio based on laboratory measurement and reanalysis of H₂ spectra’, *Phys. Rev. Lett.* **96**, 151101, 2006.
- [55] A. COC *et al.*, ‘Coupled variations of fundamental couplings and primordial nucleosynthesis’, *Phys. Rev. D* **76**, 023511, 2007.
- [56] T. DAMOUR and A. POLYAKOV, ‘The string dilaton and a least coupling principle’, *Nucl. Phys. B* **423**, 532, 1994.
- [57] M. LACHÈZE-REY and J.-P. LUMINET, ‘Cosmic topology’, *Phys. Rep.* **254**, 135, 1995.
- [58] J. LEVIN, ‘Topology and the microwave background’, *Phys. Rep.* **365**, 251, 2002.
- [59] A. LINDE, ‘Creation of a compact topologically nontrivial inflationary Universe’, *JCAP* **0410**, 004, 2004.
- [60] R. BRANDENBERGER and C. VAFA, ‘Superstring in the early Universe’, *Nucl. Phys. B* **316**, 391, 1989.
- [61] Y.B. ZELDOVICH and A.A. STAROBINSKY, ‘Quantum creation of a Universe with nontrivial topology’, *Sov. Astron. Lett.* **10**, 15, 1984.
- [62] A. RIAZUELO *et al.*, ‘Cosmic microwave background anisotropies in multiconnected flat spaces’, *Phys. Rev. D* **69**, 103518, 2004.
- [63] E. GAUSMANN *et al.*, ‘Topological lensing in spherical spaces’, *Class. Quant. Grav.* **18**, 5155, 2001.

- [64] SnapPea: <http://www.geometrygames.org/SnapPea/>.
- [65] R. LEHOUCQ, J.-P. UZAN and J. WEEKS, ‘Eigenmodes of lens and prism spaces’, *Kodai Math. J.* **26**, 119, 2003.
- [66] N.J. CORNISH, D. SPERGEL and G. STARKMANN, ‘Circles in the sky: finding topology with the microwave background radiation’, *Class. Quant. Grav.* **15**, 2657, 1998.
- [67] N. CORNISH *et al.*, ‘Constraining the topology of the Universe’, *Phys. Rev. Lett.* **92**, 201302, 2004.

13

Advanced topics

Cosmology is probably the only area where the effects of very high energy physics can have a strong impact, and even be potentially detectable. To date, the most promising theory is superstring theory that implies, amongst other things, the existence of extra dimensions.

We start by reviewing the consequences of extra dimensions for ordinary physics by focusing on Kaluza–Klein theory in Section 13.1, before describing the main aspects of string theory in Section 13.2. We then address two important themes motivated from the phenomenology of this theory, namely braneworld models in Section 13.3, and the approach of the initial singularity and bouncing universes in Section 13.4, including the pre-Big Bang scenario.

It should be noted that only recently [1] was a mechanism developed to stabilize the vacuum in string theory. As a result, most models discussed before did suffer from instabilities which render them inconsistent. However, they still can serve pedagogical purposes. We strongly advise, however, Refs. [2–4] for further, research-oriented, reading.

13.1 Extra dimensions: Kaluza–Klein theory

The space we live in is three-dimensional, and one may wonder why this is the case. This question goes back at least to Kepler (whose answer relied on the Trinity), but it was not before the beginning of the twentieth century that scientific answers could be proposed, in particular with the works of Nordström, and then with those of Kaluza and Klein. Postulating the existence of a fifth dimension, they found that, contrary to naive expectations, this hypothesis can be made compatible with the laws of physics as we know them. Considering gravity only, *a priori* the only theory capable of treating space-time geometry, they recovered general relativity in the usual four-dimensional space-time, together with a vector field that we can try to identify with the vector potential of electromagnetism and a scalar field, that would be responsible for a new interaction.

13.1.1 General relativity in D dimensions

The theory of general relativity is not specific to four-dimensional space-times, and nothing presented in Chapter 1 depends explicitly on the dimension. We can extend the entire theory to D dimensions at no additional cost, although we restrict attention to extra spatial dimensions: introducing new time dimensions is theoretically possible, but rapidly leads to twin paradoxes and to the existence of closed time-like curves, so

we shall not consider this possibility further. We will now consider an empty Universe of arbitrary dimension, so that the right-hand side of Einstein equations vanishes. The theory is thus described by the Einstein–Hilbert action in D dimensions

$$S = \frac{1}{2\kappa_D} \int R \sqrt{|g|} d^D x, \quad (13.1)$$

where g_{AB} is the metric in D dimensions, with signature $(-, +, \dots, +)$; R is the Ricci scalar associated with this metric and κ_D is the D -dimensional generalization of $\kappa = 8\pi G_N$, namely

$$\kappa_D = (D-2)\Omega^{[D-2]}G_D \quad \text{with} \quad \Omega^{[D-1]} = \frac{2\pi^{D/2}}{\Gamma(D/2)}, \quad (13.2)$$

$\Omega^{[D-1]}$ being the surface of a $(D-1)$ -dimensional sphere. In agreement with the rest of the book, we shall denote the actual 4-dimensional Newton constant by its usual symbol, namely we will set $G_4 \equiv G_N$.

The field equations of this theory are those derived in Chapter 1 for the vacuum, merely generalized to D dimensions, namely

$$G_{AB} = 0 \iff R_{AB} = 0, \quad (13.3)$$

with the indices $A, B = 0, \dots, D-1$ and G_{AB} the Einstein tensor built out of the D -dimensional metric g_{AB} .

13.1.1.1 4 + 1 decomposition

We now restrict ourselves to the specific case $D = 5$,

$$S = \frac{1}{12\pi^2 G_5} \int \bar{R} \sqrt{|\bar{g}|} d^5 x, \quad (13.4)$$

where we denote by a bar quantities in 5 dimensions to differentiate them from the analogous quantities with no bar in 4 dimensions. The aim is to determine the independent components of the metric, which are 15 in five dimensions.

Just as we perform an SVT decomposition for the cosmological perturbations into scalar, vector and tensor modes, we can proceed here with a decomposition into a symmetric tensor part $g_{\mu\nu}$, which will essentially reduce to the four-dimensional metric, with 10 independent components, a vector part, A_α , with four components that we wish to identify as the electromagnetic field, and finally a scalar field, ϕ , to complete the counting of the number of degrees of freedom ($15 = 10 + 4 + 1$). The total metric is thus written in the form

$$\bar{g}_{AB} = \begin{pmatrix} g_{\mu\nu} + \frac{1}{M^2} \phi^2 A_\mu A_\nu & \frac{1}{M} \phi^2 A_\mu \\ \frac{1}{M} \phi^2 A_\nu & \phi^2 \end{pmatrix}, \quad (13.5)$$

where the different components depend a priori both on the usual space-time coordinates x^α and the coordinate in the extra dimension y . The constant M has dimensions of mass, so that A_α also has dimensions of mass, whereas the scalar field ϕ is

here dimensionless. Finally, while capital latin indices vary in the entire 5-dimensional space-time, $A, B = 0, \dots, 4$, greek indices span the 4-dimensional space-time, namely $\mu, \nu = 0, \dots, 3$.

13.1.1.2 Cylinder conditions

A crucial point of Kaluza theory is what is called the cylinder condition. This condition, whose meaning was understood later by Klein, defines compactification on a circle. It can be formulated as follows: we assume that nothing can depend on the fifth dimension, in other words, that the coordinate y can be ignored, i.e. $\partial_y = 0$ for whatever quantity it acts upon. This is a very strong condition, which will be justified in the following, once the fifth dimension is compactified.¹ It simply implies that the scale of compactification, i.e. the radius of the fifth dimension, is very small compared to all the other characteristic lengths. This implies that to be sensitive to variations in the fifth dimension, we should be dealing with very high energies.

13.1.2 Projected equations in four dimensions

The determinant of the metric (13.5) is

$$|\bar{g}| = \phi^2 |g|, \quad \text{with} \quad g = \det(g_{\mu\nu}),$$

giving the inverse metric

$$\bar{g}^{AB} = \begin{pmatrix} g^{\mu\nu} & -\frac{1}{M} A^\mu \\ -\frac{1}{M} A^\nu & \phi^{-2} + \frac{1}{M^2} A_\alpha A^\alpha \end{pmatrix}.$$

Moreover, using the cylinder condition and the definition (1.35) of the Christoffel symbols, we find

$$\begin{aligned} \bar{\Gamma}^4_{44} &= \frac{1}{2M} A^\mu \partial_\mu \phi^2, \\ \bar{\Gamma}^4_{4\mu} &= \frac{1}{2\phi^2} \partial_\mu \phi^2 + \frac{1}{2M^2} A^\alpha (A_\mu \partial_\alpha \phi^2 - \phi^2 F_{\mu\alpha}), \\ \bar{\Gamma}^4_{\mu\nu} &= \frac{1}{2M\phi^2} [\partial_\mu (\phi^2 A_\nu) + \partial_\nu (\phi^2 A_\mu)] - \frac{1}{M} A_\beta \bar{\Gamma}^\beta_{\mu\nu} \\ &\quad - \frac{1}{2M^3} A^\alpha [\phi^2 (A_\nu F_{\mu\alpha} + A_\mu F_{\nu\alpha}) - A_\mu A_\nu \partial_\alpha \phi^2], \\ \bar{\Gamma}^\alpha_{44} &= -\frac{1}{2} \partial^\alpha \phi^2, \\ \bar{\Gamma}^\alpha_{4\mu} &= \frac{1}{2M} (\phi^2 F_\mu^\alpha - A_\mu \partial^\alpha \phi^2), \end{aligned}$$

¹The subspace formed by some dimensions can be compact, in the topological sense. In this case, the dimensions in question are said to be themselves compact.

$$\bar{\Gamma}_{\mu\nu}^{\alpha} = \Gamma_{\mu\nu}^{\alpha} + \frac{1}{2M^2} [\phi^2 (A_{\nu}F_{\mu}^{\alpha} + A_{\mu}F_{\nu}^{\alpha}) - A_{\mu}A_{\nu}\partial^{\alpha}\phi^2],$$

where we set $F_{\mu\nu} = \partial_{\mu}A_{\nu} - \partial_{\nu}A_{\mu}$.

Let us move on and consider the Ricci tensor

$$\bar{R}_{AB} = \partial_C\bar{\Gamma}_{AB}^C - \partial_B\bar{\Gamma}_{AC}^C + \bar{\Gamma}_{EC}^C\bar{\Gamma}_{AB}^E - \bar{\Gamma}_{EB}^C\bar{\Gamma}_{AC}^E,$$

which is vanishing by virtue of the Einstein equations (13.3). Given the definition of the d'Alembertian $\square\phi$, the component

$$\bar{R}_{44} = \frac{1}{4M^2}\phi^4 F_{\alpha\beta}F^{\alpha\beta} - \phi(\partial_{\alpha}\partial^{\alpha} + \Gamma_{\beta\alpha}^{\alpha}\partial^{\beta})\phi, \quad (13.6)$$

leads to the equation of motion

$$\square\phi = \frac{\phi^3}{4M^2}F_{\alpha\beta}F^{\alpha\beta}. \quad (13.7)$$

This is a Klein–Gordon equation for a massless scalar field ϕ coupled to an electromagnetic field. The mixed component, $\bar{R}_{4\mu}$, on the other hand can be written as

$$\bar{R}_{4\mu} = \frac{1}{M}\left[\bar{R}_{44}A_{\mu} + \frac{1}{2}\phi^2\left(\nabla^{\alpha}F_{\mu\alpha} + 3\frac{\partial^{\alpha}\phi}{\phi}F_{\mu\alpha}\right)\right], \quad (13.8)$$

which implies, taking into account (13.7),

$$\nabla^{\alpha}F_{\mu\alpha} = -3\frac{\partial^{\alpha}\phi}{\phi}F_{\mu\alpha}. \quad (13.9)$$

This equation reduces to the Maxwell equations in vacuum provided that the scalar field is constant. Finally, the purely 4-dimensional part reads

$$\bar{R}_{\mu\nu} = R_{\mu\nu} - \frac{1}{\phi}\nabla_{\mu}\nabla_{\nu}\phi - \frac{\phi^2}{2M^2}F_{\mu}^{\alpha}F_{\nu\alpha} + \frac{1}{M}(A_{\mu}\bar{R}_{4\nu} + A_{\nu}\bar{R}_{4\mu}) - \frac{1}{M^2}\bar{R}_{44}A_{\mu}A_{\nu}, \quad (13.10)$$

where the two last terms cancel as soon as (13.7) and (13.9) are satisfied. Einstein equations (13.3) can be used to rewrite (13.10) in the form

$$R_{\mu\nu} = \frac{1}{\phi}\nabla_{\mu}\nabla_{\nu}\phi + \frac{\phi^2}{2M^2}F_{\mu}^{\alpha}F_{\nu\alpha}.$$

Using (13.7) to express part of the term in $\square\phi$ in R , we obtain the modified four-dimensional Einstein equations

$$G_{\mu\nu} = \frac{\phi^2}{2M^2}T_{\mu\nu}^{(\text{elec})} + \frac{1}{\phi}(\nabla_{\mu}\nabla_{\nu}\phi - g_{\mu\nu}\square\phi), \quad (13.11)$$

where $T_{\mu\nu}^{(\text{elec})} = F_{\mu}^{\alpha}F_{\nu\alpha} - \frac{1}{4}F_{\alpha\beta}F^{\alpha\beta}$ is the energy-momentum tensor of the electromagnetic field [see (1.84)].

All these equations can also be obtained from the variation of the action (13.4), which can now be rewritten in the form of a four-dimensional integral

$$S = \frac{1}{16\pi G_N} \int d^4x \sqrt{-g}\phi \left(R - \frac{\phi^2}{4M^2} F_{\alpha\beta} F^{\alpha\beta} \right), \quad (13.12)$$

where we have set

$$G_N = \frac{3\pi \bar{G}_5}{4V_{(5)}},$$

and factored out the finite volume of the fifth dimension, $V_{(5)} = \int dy$. In this framework, the Newton constant therefore finds a geometric meaning. We will see later that this relation has been proposed to provide a framework explaining the hierarchy problem, that is why the scale of gravity is so much higher than all the other energy scales of particle physics. We, furthermore, see that the action (13.12) does not include any term involving ϕ alone. This is due to the fact that the corresponding term in R is $\phi^{-1}\square\phi$ that, when taking into account the determinant of the metric, becomes proportional to $\square\phi$. This is therefore a total derivative that does not contribute to the dynamics.

13.1.3 Dilaton and Einstein–Maxwell theory

Many generic consequences of string theory can be drawn from the previous equations. First, if the electromagnetic field A_μ is set to zero, which is a solution of the field equations, then we recover a scalar-tensor theory of the type (10.1) with no potential in the Brans–Dicke parameterization (10.2), provided we set $\varphi = \phi^{-1}$ and $\omega(\varphi) = 0$. It is then sufficient to minimally couple this action to matter to recover exactly all the results of Section 10.1 from Chapter 10. In the framework of string theory, we often write the scalar field ϕ in the form

$$\phi = \exp \left(\frac{4\sqrt{\pi}}{\sqrt{3}M_p} \varphi \right). \quad (13.13)$$

The new scalar field φ , has the correct dimension of mass and is called the *dilaton*. In this context, the Jordan frame is called the string frame.

The configuration for which ϕ is constant is not a solution of the field equations. Therefore, it is not possible to recover the Einstein–Maxwell theory describing gravity and electromagnetism in a unified way, despite the initial hope of Einstein and Kaluza. Having said that, if we require that the first term of the right-hand side of (13.11) is exactly the same as that in general relativity, we find that we should define an effective gravitational constant G_{eff} , which will depend in particular on time in a cosmological framework, where ϕ is a function of time

$$G_{\text{eff}} = \frac{\phi^2}{16\pi M^2};$$

interestingly enough, this constant is independent of the five-dimensional gravitational constant G_5 . Note that, as explained in detail in Chapter 10 [(10.18) and (10.19)], this

constant is not the one we would measure in a Cavendish kind of experiment: for such a measurement, one would need to consider real matter, which we did not include yet. Defining the electric and magnetic fields, \mathbf{E} and \mathbf{B} , respectively, associated to the vector potential A_μ , (13.7) can be rewritten in the form

$$\square\phi \sim 16\pi^{3/2} \left(\frac{M}{M_p} \right) \frac{\mathbf{E}^2 - \mathbf{B}^2}{M_p^2}, \quad (13.14)$$

where a ϕ dependence should appear in the right-hand side in the Planck mass; we have here simply replaced it by the numerical value currently measured for this mass, i.e. we have set $G_{\text{eff}} = M_p^{-2}$ and reported this value for ϕ , namely $\phi \sim 4M_p\sqrt{p}/M_p$, into (13.7). We then note that for most usual situations for which the electric and magnetic fields have a very weak amplitude compared to the Planck mass currently measured, the right-hand term is very small, and all the more so that the mass scale of the fifth dimension is small compared to the Planck mass. Therefore, treating ϕ as constant can be a good approximation, and one recovers the Einstein–Maxwell theory. Note, however, that such a dilaton will imply a variation of the fine-structure constant and is generally excluded by tests of the equivalence principle [5] (see Section 12.3 of Chapter 12).

13.1.4 Einstein frame

There is a subtlety [6] concerning the action (13.12) stemming from the fact that it does not seem to contain an explicit kinetic term for the scalar field ϕ , although it is sometimes written, as a shorthand, in review articles on this topic [7]. However, because ϕ is subject to (13.7), it is indeed a dynamical variable. This apparent paradox is easily resolved by noting that the dynamics is, in fact, present, as explained in Chapter 10. Again, to see this, let us look at the variation with respect to ϕ of (13.12). This gives a relation between R , ϕ and $F_{\alpha\beta}F^{\alpha\beta}$, whereas that with respect to the metric gives $R_{\mu\nu}$ as a function of ϕ . Taking the trace of this last equation and applying the constraint between ϕ and R , one precisely recovers (13.7).

This point becomes immediately transparent as soon as one diagonalizes the degrees of freedom by switching to the Einstein frame. In D dimensions, under the ‘conformal’ transformation $\bar{g}_{AB}^* = \Omega^2 \bar{g}_{AB}$, we have $\sqrt{|\bar{g}^*|} = \Omega^D \sqrt{|\bar{g}|}$, and (10.76) indicates that the curvature term is modified according to $\sqrt{|\bar{g}^*|}\bar{R}^* = \Omega^{D-2} \sqrt{|\bar{g}|}\bar{R} + \dots$. Compactifying to $D-1$ dimensions with $(D-1)+1$ decomposition similar to (13.5) for \bar{g}_{AB} , we have that $\sqrt{|\bar{g}^*|} = \phi \sqrt{|g|}$. Therefore, the pure Einstein–Hilbert action for the ‘starred’ variables has a geometric term that is $\sqrt{|g|}\phi\Omega^{D-2}\bar{R} \rightarrow \sqrt{|g|}\phi\Omega^{D-2}R$, where the last step stems from (13.10).

Restricting now again to $D=4$ and demanding that general relativity holds in these 4 dimensions, we see that we must set $\Omega = \phi^{1/(2-D)} = \phi^{-1/2}$. Gathering the previous expressions, we then find

$$\begin{aligned}
S &= \frac{1}{12\pi^2 \bar{G}_5} \int d^5x \sqrt{-\bar{g}^*} \bar{R}^* \\
&= \frac{1}{16\pi G_N} \int d^4x \sqrt{-g} \left(R - \frac{\phi^2}{4M^2} F_{\alpha\beta} F^{\alpha\beta} - \frac{3}{2} g^{\mu\nu} \frac{\nabla_\mu \phi \nabla_\nu \phi}{\phi^2} \right), \\
&= \int d^4x \sqrt{-g} \left\{ \frac{1}{16\pi G_N} \left[R - \frac{\phi^2(\varphi)}{4M^2} F_{\alpha\beta} F^{\alpha\beta} \right] - \frac{1}{2} \nabla_\mu \varphi \nabla^\mu \varphi \right\}, \quad (13.15)
\end{aligned}$$

where we have used the definition (13.13) for φ . It is then obvious that the field behaves as a canonical scalar field.

13.1.5 Compactification

The fundamental modification brought by Klein to what precedes was to understand that the cylinder condition was justified as long as we consider the fifth dimension to be topologically compact with the topology of a circle. In this case, all the fields that are defined in this space, i.e. the four-dimensional metric $g_{\mu\nu}$, the vector potential A_α and the dilaton ϕ , and any additional matter fields that the theory should describe, are periodic functions of the extra dimension and can therefore be expanded into Fourier modes. The radius R of this dimension then turns out to be naturally $R \sim M^{-1}$, and we understand why we do not see variations in the circle: for a large enough M , its radius is too small to be observable. To be sensitive to the fifth dimension, the energies involved must be comparable to M , so that all the effects will be unobservable if the radius is small enough.

Let us illustrate this in the example of a real massless scalar field Σ in five dimensions. Its action in the Einstein frame is simply

$$S_\Sigma = -\frac{1}{2} \int d^5x \sqrt{|\bar{g}^*|} \bar{g}^{*AB} \partial_A \Sigma \partial_B \Sigma, \quad (13.16)$$

and we expand this field in Fourier series, still denoting by y the coordinate in the fifth dimension. We thus write

$$\Sigma(x_\mu, y) = \sum_{n=-\infty}^{+\infty} \sigma_n(x_\mu) e^{inMy}, \quad \text{with} \quad \sigma_{-n} = \sigma_n^* \quad (\Sigma \in \mathbb{R}), \quad (13.17)$$

and then expand (13.16) explicitly assuming that $\partial_y \phi = 0 = \partial_y A_\mu$, i.e. we are only interested in the zero modes of ϕ and A_μ . Using the one-dimensional equation (B.57) to integrate along y , i.e. using $\int e^{i(n+p)My} dy = 2\pi\delta(p+n)/M$, we then find

$$S_\Sigma = -\frac{2\pi}{M} \int d^4x \sqrt{-g} \left\{ \frac{1}{2} \partial_\mu \sigma_0 \partial^\mu \sigma_0 + \sum_{n=1}^{+\infty} \left[|(\partial_\mu - inA_\mu) \sigma_n|^2 + \frac{n^2 M^2}{\phi^2} |\sigma_n|^2 \right] \right\}. \quad (13.18)$$

It remains only to change the normalization of the scalar fields by $\sigma_n = \sqrt{M/4\pi} \tilde{\sigma}_n$ to obtain the action for a real massless canonical field $\tilde{\sigma}_0$ and an infinity of complex canonical fields $\tilde{\sigma}_n$ of masses $m_n = nM^{3/2}/(\phi\sqrt{2\pi})$, which depend explicitly on the value of the dilaton at the position and time for which we evaluate them.

Considering now the ‘electromagnetic’ term in the action (13.15), we see that the dilaton can be absorbed into a new normalization of the gauge field by setting $A_\mu^{\text{elec}} = \phi A_\mu$, from where we exactly recover the Maxwell term and the coupling corrections between the dilaton and the electromagnetic field. From the point of view of the scalar fields in (13.18), this leads to kinetic terms of the form $(\partial_\mu - inA_\mu^{\text{elec}}/\phi)\sigma_n$, in other words charged fields with charges $e_n = n/\phi$. It is interesting to note that not only are these charges automatically quantized, but here again they depend on the value of the dilaton. As a consequence, assuming that the first Kaluza–Klein mode has unit charge, the electromagnetic coupling constant is

$$\alpha_{\text{EM}} = \frac{e_1^2}{4\pi} = \frac{1}{4\pi} \exp\left(-\frac{4\pi}{\sqrt{3}M_p}\varphi\right), \quad (13.19)$$

and we get the right value $\alpha_{\text{EM}}^{-1} \sim 137$, provided that the value of the dilaton is of order $\phi \sim 10^{-2}M_p$, compatible with the string scale. If the dilaton has dynamics this relation could induce a variation of the fine structure constant, which is highly constrained (Section 12.3 of Chapter 12).

13.2 A few words on string theory

String theory is based on the principle that the fundamental objects are no longer point particles but one-dimensional objects, strings, with a very small characteristic length compared to the characteristic scales involved in the experiments realized so far (and vanishing thickness). To a good approximation, they therefore appear point-like in all these experiments. One can find many excellent reference works [8] on this fashionable subject, and also many regularly updated websites [9], and so we shall provide below merely a summary of the most important features for cosmology.

13.2.1 From particles to strings

13.2.1.1 Classical action

The worldline of a particle is given by the least action principle, according to which the action should be minimized, namely, in Minkowski space-time,

$$S_0 = -m \int d\tau \sqrt{-\eta_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau}} = -m \int \sqrt{1 - v^2} dt,$$

where τ is the particle proper time, m its mass and $v \equiv dx/dt$. This action can be generalized to the case where the fundamental object is not a point-like particle, of zero spatial dimension, but a ‘surface’ of spatial dimension p , that we call a p -brane. If T_p is the uniform tension of this p -brane, the action above is generalized into

$$S_p = -T_p \int \sqrt{-\det \left(\eta_{\mu\nu} \frac{\partial x^\mu}{\partial \xi^a} \frac{\partial x^\nu}{\partial \xi^b} \right)} d^{p+1}\xi, \quad (13.20)$$

where the ξ^a , $a = 1, \dots, p+1$ are coordinates along the p -brane, and the determinant is that of the $(p+1) \times (p+1)$ matrix whose lines and rows are labelled by a and b .

In the case where $p = 0$, we identify the tension to the particle mass, $T_0 = m$, and we recover the action of a point-like particle. Let us now consider, for definiteness, the specific case of a string, in other words a 1-brane, of tension $T_1 = T$. Equation (13.20) is rewritten in the form

$$\mathcal{S}_1 = -T \int d\tau d\sigma \sqrt{\dot{x}^2 x'^2 - (\dot{x} \cdot x')^2}, \quad \dot{x}^\mu \equiv \frac{\partial x^\mu}{\partial \tau}, \quad x'^\mu \equiv \frac{\partial x^\mu}{\partial \sigma}, \quad (13.21)$$

where $\tau = \xi^0$ and $\sigma = \xi^1$ form a set of, respectively, time-like and space-like coordinates along the string. The action (13.21), which is the Nambu–Goto action, is more easily written in the form

$$\mathcal{S}_{NG} = -T \int d^2\xi \sqrt{-\gamma}, \quad \text{with} \quad \gamma_{ab} \equiv \eta_{\mu\nu} \frac{\partial x^\mu}{\partial \xi^a} \frac{\partial x^\nu}{\partial \xi^b}, \quad (13.22)$$

where γ_{ab} is the induced metric on the string worldsheet (see Section 8.4.5 of Chapter 8).

Another way to look at a string dynamics is to consider the two-dimensional metric elements along the string as independent variables, that will be determined later as solutions of the equations of motion. In this case, initially we no longer have the definition of (13.22), and the action reduces to the Polyakov action

$$\mathcal{S}_P = -\frac{T}{2} \int d^2\xi \sqrt{-h} h^{ab} \eta_{\mu\nu} \frac{\partial x^\mu}{\partial \xi^a} \frac{\partial x^\nu}{\partial \xi^b}. \quad (13.23)$$

We then have, in addition to the initial equations of motion, a constraint that allows us to fix the string metric. This constraint is obtained by varying the action (13.23) with respect to h_{ab} , and one finds

$$T_{ab} \equiv \eta_{\mu\nu} \left(\frac{\partial x^\mu}{\partial \xi^a} \frac{\partial x^\nu}{\partial \xi^b} - \frac{1}{2} h_{ab} h^{cd} \frac{\partial x^\mu}{\partial \xi^c} \frac{\partial x^\nu}{\partial \xi^d} \right) = 0. \quad (13.24)$$

We often write the string tension T in the form

$$T = \frac{1}{2\pi\ell_s^2} = \frac{1}{2\pi\alpha'},$$

(13.25)

where α' is called the universal Regge slope, and ℓ_s the characteristic length of the fundamental strings.

13.2.1.2 Symmetries and dimensions

The actions describing the string trajectories are parameterization invariant, i.e. invariant under general coordinate transformations on the worldsheet, just as in general relativity. In other words, replacing the coordinate system $\{\xi^a\}$ by any other one, $\{\tilde{\xi}^b\}$, arbitrary functions of ξ^a , we shall recover the same action, expressed in terms of $\tilde{\xi}^b$. For a string, h_{ab} has only three independent components, so that one can fix

two of them and finally express everything in terms of a unique degree of freedom. A convenient parameterization is that for which we choose the conformally flat gauge

$$h_{ab} = e^{\phi(\sigma, \tau)} \eta_{ab}, \quad (13.26)$$

where $\eta_{ab} = \text{diag}(-1, 1)$ is the two dimensional Minkowski metric.

Moreover, in the Polyakov form, the action has a complementary so-called conformal invariance: (13.23) is unchanged if we replace the elements of the string metric h_{ab} by $f(\sigma, \tau)h_{ab}$, i.e. if we perform a conformal transformation of the 2-dimensional metric, while keeping the string coordinates x^μ fixed. This symmetry is only present for the case of a string and we thus expect in this case that the function ϕ of (13.26) to be no longer physically relevant. Classically, this is the case since if we evaluate the Polyakov action in a conformally flat gauge, we simply find

$$S_p = -\frac{T}{2} \int d^2\xi \eta^{ab} \eta_{\mu\nu} \frac{\partial x^\mu}{\partial \xi^a} \frac{\partial x^\nu}{\partial \xi^b},$$

which describes the dynamics of D scalar fields x^μ , ($\mu = 0, \dots, D-1$) when the space in where the strings evolve is D -dimensional.

So far, the generalization of point-like particles to the string case is almost trivial. The situation changes as soon as quantum effects are taken into account. Writing the generators of the Lorentz group in D dimensions, we find that the commutation relations between these generators produce a quantum *anomaly*, so that the Lorentz invariance cannot be respected. This anomaly exists for any space-time dimension, but vanishes identically if $D = 26$. The theory based on these so-called ‘bosonic strings’ therefore appears to be physically consistent only in 26 dimensions.

13.2.2 Superstrings

String theory predicts that space-time should have 26 dimensions. This is the first time that the number of dimensions is not fixed arbitrarily but is imposed by the requirement of mathematical consistency of the theory. From this point of view, this is a success of string theory. Unfortunately, this is only a partial success since the value $D = 26$ does not correspond to the observed value $D = 4$. Furthermore, we find that such a theory is unstable since some excitation modes of the strings have a negative squared mass: they are tachyons.

One way to solve this stability problem is to modify the previous theory to include fermions, and not only describe the bosonic states (the x^μ). To do so, and in order for the theory to be meaningful, it turns out that one also must impose a supersymmetry between fermions and bosons. This stems from the fact that the zero-point energies of bosons and fermions are exactly equal up to their signs, so that their sum vanishes identically (Chapter 10), and the tachyonic mode disappears from the excitation spectrum of the theory. Its two-dimensional effective action is given by the supersymmetric version of the Polyakov action, hence containing D bosonic fields and as many fermionic ones, namely

$$S = -\frac{1}{4\pi\alpha'} \int d^2\xi \eta_{\mu\nu} \left(\eta^{ab} \frac{\partial x^\mu}{\partial \xi^a} \frac{\partial x^\nu}{\partial \xi^b} + i\bar{\psi}^\mu \gamma^a \frac{\partial}{\partial \xi^a} \psi^\nu \right), \quad (13.27)$$

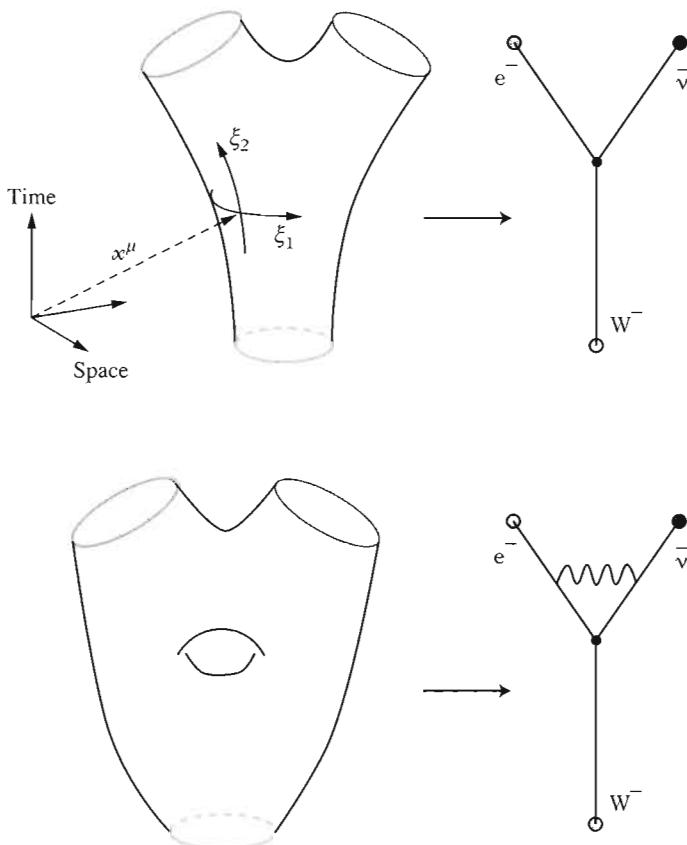


Fig. 13.1 The string worldsheets can be understood as enlargements of the point-like particle worldlines, loops in the interactions then appear as holes in the topology of the surface described by the string trajectories. This is how ordinary particle physics is presumed to be a low-energy approximation of string theory. We also see how one of the problems of field theory, namely the existence of interaction points (vertices) in the interactions of point-like particles that produce technical difficulties (divergences to renormalize) is not present in a string description.

where the degrees of freedom ψ^μ are fermionic, and the matrices γ^a are 2×2 matrices satisfying $\{\gamma^a, \gamma^b\} = -2\eta^{ab}$.

The resulting theory is now called a *superstring* theory, which it is found to be only physically consistent in $D = 10$ dimensions (by the same Lorentz anomaly token). This is slightly better, but we understand that this still requires six compact extra dimensions. The Kaluza–Klein effects described earlier are then expected as a generic consequence of superstring theories.

13.2.2.1 Quantum gravity

A generic consequence of string and superstring theories is related to the existence of a specific state in the set of the possible excitations of the string. One finds that in all these theories, there is a massless spin-2 degree of freedom that one naturally identifies with the graviton, as well as a scalar. In other words, string theories all include a scalar-tensor theory of gravitation, which is, in addition, consistent at the quantum level. This explains why these theories are considered to be good candidates for quantum gravity. This is indeed, to date, the only way to compute without any divergences the scattering interaction between two gravitons, a process that still belongs to the widely unknown area of quantum gravity. This calculation is made possible by the fact that the interaction is realized through the exchange of strings, which, as shown in Fig. 13.1, a priori leads to less divergences to be renormalized than point-particle interactions.

13.2.2.2 Five (plus one) possible theories

Strings can exist in different possible forms, and this leads to different possible theories. Actually, only five of them are consistent, to which one should add supergravity in 11 dimensions for technical reasons (only in this case can supergravity be made consistent at the quantum level). One thus obtains a total of six theories.

It was shown that there exist relationships between these possible string theories, and that therefore they are not independent. It was then conjectured that they could originate from a single theory, which was called \mathcal{M} -theory: one expects that the five 10-dimensional string theories as well as 11-dimension supergravity are merely different limits, in different regimes, of this hypothetical \mathcal{M} -theory. We usually summarize this set of theories with the star-shaped diagram of Fig. 13.2.

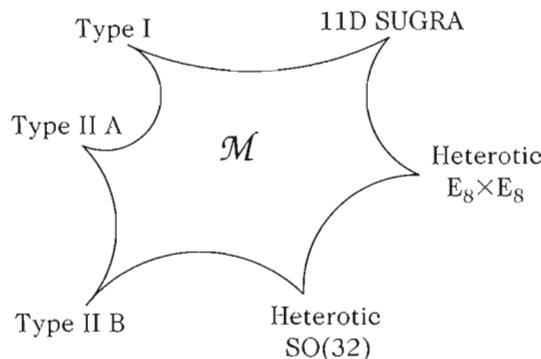


Fig. 13.2 \mathcal{M} -theory is assumed to be the ‘ultimate’ theory of fundamental interactions, from which the 5 mathematically consistent string theories, in 10 dimensions, as well as supergravity in 11 dimensions are derived. There are indeed sufficient duality relations between these different theories to argue that they should be unified into a unique one, of which they would occupy different limits in the parameter space.

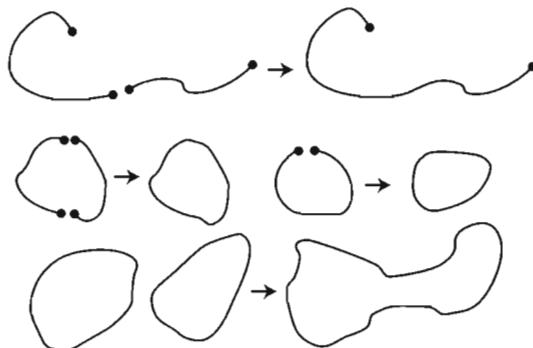


Fig. 13.3 Open and closed strings interactions. We see that open strings can produce loops, whereas closed string remain closed.

13.2.3 Open and closed strings

Strings can be of two types, either open, with free extremities, or closed in the form of loops. Interactions between open string extremities are such that when two extremities meet, they join to form a longer string. We then see that in some cases, illustrated in Fig. 13.3, open strings can form closed strings, so that an open string theory must also be able to describe closed strings, whereas a string theory containing only closed strings can be consistent by itself. It was found that there is only one possible theory of open supersymmetric strings, called type I, and two closed string theories, called types IIA and IIB. Furthermore, there are two closed string theories in which supersymmetric and non-supersymmetric states are present; those have been called ‘heterotic’ strings. In other words, among the five mathematically consistent string theories, only one contains open strings, while the other four describe loops exclusively.

13.2.3.1 Type I

The lowest-energy excitations of string theories can be described by a field theory with a fixed particle spectrum, in a ten-dimensional space-time. For open strings of type I theory, one first obtains the graviton, together with a set of gauge fields that can be shown to form a representation of the $SO(32)$ symmetry, as well as fermions, transforming in a non-trivial way with respect to transformations of $SO(32)$. These fermions have definite chirality, so that the type I theory does not conserve parity, as it should in order to reproduce the properties of the electroweak standard model. All these particles are massless, which can be interpreted as a very high-energy effect, knowing that at low energies, just as for theories with spontaneous symmetry breaking, we expect to recover non-vanishing masses.

One can raise the question of the boundary conditions for the string extremities. Requiring that Lorentz invariance is respected in D dimensions, these boundary conditions should then be compatible with the fact that these string extremities can be anywhere, since the restriction to some places would produce preferred regions. The conditions that respect the Lorentz invariance are called Neumann conditions, following the standard nomenclature of differential equations. When the string extremities

are fixed in some specific regions, we then apply the so-called Dirichlet boundary conditions. Despite breaking Lorentz invariance, these conditions are quite reasonable as soon as we consider particle states, since only the vacuum is supposed to respect this invariance.

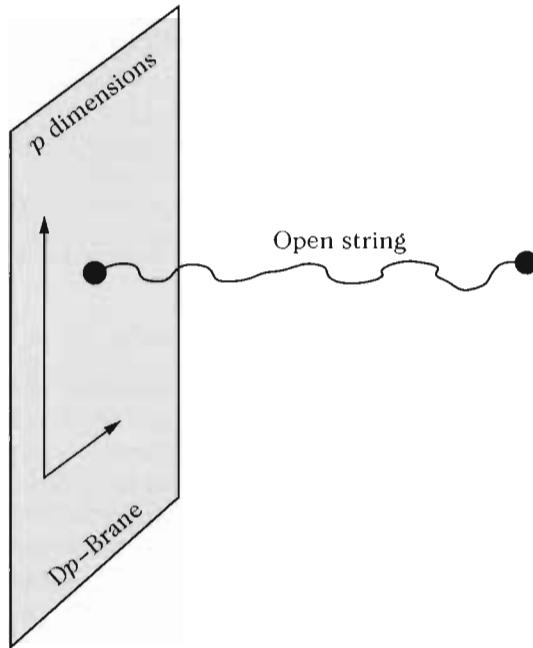


Fig. 13.4 D_p-brane formed by the set of points of an open string extremities with $9 - p$ coordinates satisfying Dirichlet boundary conditions.

The equations of motion obtained by varying the action (13.22) are simply

$$\ddot{x}^\mu - x''^\mu = 0, \quad (13.28)$$

and the constraint (13.24) then becomes

$$\dot{x} \cdot x' = 0, \quad \dot{x}^2 + x'^2 = 0 \quad \Rightarrow \quad (\dot{x} \pm x')^2 = 0. \quad (13.29)$$

For a string with space-like parameter σ varying by convention between 0 and π , the Neumann and Dirichlet conditions become

$$\text{Neumann condition: } \frac{\partial x^\mu}{\partial \sigma} = 0 \quad \text{for} \quad \sigma = 0, \pi, \quad (13.30)$$

and

$$\text{Dirichlet condition: } \frac{\partial x^\mu}{\partial \tau} = 0 \quad \text{for} \quad \sigma = 0, \pi. \quad (13.31)$$

We see that this last condition can be integrated, defining the precise location of the space where the extremity is located.

One can, for instance, impose that the string extremity is fixed at one point in the nine-dimensional space. In this case the point in question could represent a point-like particle, different from a string excitation state. It will then be called a D-particle (for Dirichlet). One can also imagine mixed conditions between Neumann and Dirichlet, say, for instance, p Neumann conditions and $9 - p$ Dirichlet conditions. In this case, we naturally see a new category of objects in the spectrum of the theory, having p spatial dimensions, and called D p -branes.

13.2.3.2 Type IIA and IIB

The situation is very different in the case of closed supersymmetric string theories. Besides the graviton, these theories do not contain any gauge boson, so the fermions described by these theories are a priori not charged. Gauge bosons appear as soon as we compactify, and fermions then acquire charges. In the type IIA theory, parity is conserved, which is not the case of the type IIB. Finally, these theories also describe tensor gauge fields, which are gauge fields with several indices, which generalizes the action

$$S_{\text{elec}} = -m \int_{\text{trajectory}} ds + q \int_{\text{trajectory}} A_\mu(x) dx^\mu - \frac{1}{4m_e^{D-4}} \int F_{\mu\nu} F^{\mu\nu} d^D x, \quad (13.32)$$

describing a particle of charge q coupled to electromagnetism (the D -dimensional coupling constant is dimensioned). In (13.32), the arbitrary mass, m_e appears in front of the kinetic term of the gauge field due to the number of dimensions: the case $D = 4$ for which no such arbitrary mass is present in the action is special for the gauge fields. One then finds

$$S = S_1 - \frac{1}{2} \int d\tau d\sigma B_{\mu\nu} \frac{\partial x^{[\mu}}{\partial \tau} \frac{\partial x^{\nu]}}{\partial \sigma} - \frac{1}{6\tilde{m}^{D-6}} \int H_{\mu\nu\rho} H^{\mu\nu\rho} d^D x, \quad (13.33)$$

where S_1 is the Nambu–Goto action (13.21). The antisymmetric tensor gauge field $B_{\mu\nu}$ is called the Kalb–Ramond field, and

$$H_{\mu\nu\rho} = \partial_\mu B_{\nu\rho} + \partial_\nu B_{\rho\mu} + \partial_\rho B_{\mu\nu}$$

is the equivalent of the Faraday tensor for 2-forms. The presence of the Kalb–Ramond field makes it possible for the loops to carry a conserved charge (the calculation is made by integrating its flux in a way completely similar to what is done in electromagnetism), so that the corresponding loops are stable.

13.2.3.3 Heterotic strings

Like type II strings, heterotic strings are closed. One can express the general solution of the Nambu–Goto equation of motion (13.28) in terms of modes of left and right chirality, (respectively, left and right movers)

$$x^\mu(\sigma, \tau) = x_L^\mu(\tau + \sigma) + x_R^\mu(\tau - \sigma). \quad (13.34)$$

These right and left bosonic modes, despite propagating along the same string, do not interact with one another. To make the theory completely supersymmetric, it is

sufficient to associate to each of these modes a fermionic partner propagating in the same way.

One can also associate fermionic modes only to a subset of bosonic modes, say the right movers to fix the ideas, and this is what realizes the heterotic string, immediately raising a technical difficulty. We have seen that the bosonic string was only consistent in 26 dimensions, whereas a supersymmetric string is only meaningful in 10 dimensions, so it is not obvious to know the dimension that is needed of the space for the heterotic string, which is a mixture of both, to be mathematically consistent. The solution is to take the bosonic theory, thus in 26 dimensions, and to compactify 16 of them à la Kaluza–Klein. The resulting theory is then supersymmetric in 10 dimensions, the left modes finding their origin in the 26 dimensions, introducing additional degrees of freedom: these are gauge bosons. There are only two consistent topological possibilities to compactify the 16 additional dimensions leading to two different gauge groups: $\text{SO}(32)$, as in the case of type I open strings, or $E_8 \times E_8$, which are both of rank 16, i.e. the number of compactified dimensions, and have 496 generators (see Table 2.2). Neither of these two theories respect parity.

13.2.3.4 11-dimensional SUGRA

Although strictly speaking not a string theory, we have seen that 11-dimensional supergravity appears naturally in the framework discussed above. In some regimes, discussed later (Section 13.3.1), a coupling constant can in fact depend on the distance between two 10-dimensional spaces. Only gravity, because it is supersymmetric, can propagate through the eleventh dimension, while all the other fields are restricted to one of the 10-dimensional spaces. It is interesting to note that, by coincidence, supergravity is only meaningful at the quantum level in 11 dimensions. This provides a hint that the actual \mathcal{M} -theory ought to be also in 11 dimensions, string theories then appearing 10-dimensional simply because the eleventh dimension must be so small that the string approximation is valid.

13.2.4 Dualities

All these theories, including 11-dimensional supergravity, turn out to be related to one another through duality relations, and this was the reason to assume, in the first place, the existence of a more general theory containing them all. The realization of this fact is what was called the ‘second superstring revolution’.

The first duality, called ‘S’-duality, can be described in the following way. Two string theories, A and B , with coupling constants g_s , are said to be S-dual if the predictions of one of them evaluated in a strong coupling regime are the same as that of the other theory in the weak coupling regime. In other words, for a physical quantity P that we wish to compute, then

$$P_A(g_s) = P_B\left(\frac{1}{g_s}\right),$$

where $P_{A,B}$ are the quantities computed in the theories A and B , respectively. It happens that type I theory is related to $\text{SO}(32)$ heterotic theory via this duality, and type IIB is self-dual. When the coupling increases, the other two theories, type IIA

and heterotic $E_8 \times E_8$, behave as if they were effectively in 11 dimensions. Expressing string theory in terms of Kaluza–Klein fields, it turns out that the string coupling constant is the value of the scalar part ϕ . Therefore, according to (13.13), demanding the S-duality is equivalent to imposing a $\varphi \rightarrow -\varphi$ invariance [see also (13.19)].

The second duality, called ‘T’-duality, makes the connection between the compactification of the different theories. If the A and B theories have compact dimensions of characteristic size R_A and R_B , then they are T-dual if their physical predictions are the same (the theories are equivalent) when R_A and R_B are related by

$$R_A R_B = \ell_s^2.$$

This symmetry can be illustrated via a compactification on a circle of radius R . A particle in motion around this circle sees its momentum quantized in integer multiples of $1/R$, which are simply the Kaluza–Klein modes. Moreover, a string can also wind around the circle an integer number of times. Now, the larger the circle, the greater the energy of the winding mode. We see that exchanging R and $1/R$ and replacing the winding and momentum modes leads to the same configuration. We find such T-dualities between type II and heterotic string theories.

One can conclude that all the string theories are effectively somehow dual to one another, and also dual to 11-dimensional supergravity; hence \mathcal{M} -theory.

13.2.5 Low-energy Lagrangians

Most of the relevant predictions of string theories to cosmology can be made only from the low-energy gravitational effects to describe the global evolution of the Universe, together with a few string corrections. In principle, such a regime should allow for a description of the phase close to the primordial singularity. At low energies one can reduce the complete theory, in which gravity comes out naturally from spin-2 modes, to a field theory in a given, arbitrary, background metric.

13.2.5.1 Bosonic string

Let us consider the case of the bosonic states alone, or in other words let us work in the case of the bosonic string. We hence neglect the fermions in the two-dimensional action (13.27), but we add the coupling term of the type (13.33) for the antisymmetric Kalb–Ramond tensor as well as the kinetic term for the graviton. The action (13.23) generalizes to

$$S_b = -\frac{1}{4\pi\alpha'} \int d^2\xi \sqrt{-\gamma} \left\{ (\gamma^{ab} g_{\mu\nu} + i\epsilon^{ab} B_{\mu\nu}) \frac{\partial x^\mu}{\partial \xi^a} \frac{\partial x^\nu}{\partial \xi^b} + 2\alpha' R[x(\xi)] \phi[x(\xi)] \right\}, \quad (13.35)$$

where the dilaton ϕ appears naturally as a general coupling function for the curvature dynamic term, and the background metric $g_{\mu\nu}$ is now considered as an independent degree of freedom. This last assumption implies that the kinetic terms of the bosons x^μ are now to be seen as couplings between the various metric degrees of freedom. This is similar to what happens to scalar fields in a curved background: in both cases, this derivative coupling allows a pair of bosons (scalar fields or x^μ fields) to be annihilated into a graviton $g^{\mu\nu}$.

The action (13.35) is a two-dimensional action, for which we require Weyl conformal invariance, necessary for the theory to be consistent both at the classical and quantum level. This non-trivial invariance imposes constraints that generalize (13.24). For small values of the fields $g_{\mu\nu} - \eta_{\mu\nu}$ (small departure from the Minkowski metric), $B_{\mu\nu}$ and ϕ , we can expand the constraints to different orders of approximation, to obtain the field equations of motion. These equations take the form of an expansion in powers of the constant α' . This is the origin of the term α' correction terminology often found in the literature to describe the difference between string theory and general relativity, for instance.

The constraint equations we obtain are threefold. First, we have those stemming from general relativity coupled to a dilaton, i.e. those of a scalar-tensor theory. Then, we obtain the dynamical equation of the Kalb–Ramond field, that we would have obtained by varying the D -dimensional term of (13.33). Finally, we have an equation to drive the dilaton dynamics. These equations can also be obtained from the variations of an effective field-theory action in D dimensions, namely

$$S_{\text{eff}} = \frac{1}{2\kappa_D} \int d^Dx \sqrt{-g} e^{-\phi} \left(R + g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi - \frac{1}{12} H_{\mu\nu\rho} H^{\mu\nu\rho} \right) + \mathcal{O}(\alpha'), \quad (13.36)$$

where we neglected the α' corrections and where $H_{\mu\nu\rho}$ is defined by (13.33); the constant κ does not influence the equations of motion since it can always be modified by a dilaton rescaling and is thus merely a normalization.

We can, and, in general, we do, obtain extra fields in the different versions of the theory, and these fields can play important roles in cosmology. Having said that, the three terms present in the action (13.36) appear in all models of string cosmology.

The configuration with $g_{\mu\nu} = \eta_{\mu\nu}$, $B_{\mu\nu} = 0$, and $\phi = \phi_0$ is a solution of the equations of motion, i.e. it is invariant under conformal transformations. It can be seen from (13.35) that the scalar field ϕ defines a coupling between the strings, as it appears in front of the geometric term: as in ordinary field theory, it is this kind of term that allows for strings to evolve in a non-trivial way, hence to interact. It is usual to set the string coupling constant as

$$g_s = \phi \quad \implies \quad \ln g_s \propto \varphi.$$

In field theory, however, the value of the coupling constant is fixed once and for all: taking a different value for this measurable constant amounts to considering a physically different theory. In the case at hand, one can modify the value of the coupling constant in a dynamical way, while keeping the same theory, since different values are possible as different solutions of the field equations.

13.2.5.2 Compactification

The theories considered so far still cannot be used in cosmology in their present state as they require a 10-dimensional space-time (or even 11 for supergravity). There are therefore 6 dimensions that should thus be made ‘invisible’ by a compactification similar to Kaluza–Klein theory. In other words, we will, in general, write that the total 10-dimensional space-time, \mathcal{X} , can be decomposed into $\mathcal{X} = \mathcal{M} \times \mathcal{K}$, where \mathcal{M}

is the usual four-dimensional space-time, assumed to be non-compact, and \mathcal{K} is of too small a characteristic size to be observable.

For mathematical reasons, the compact space \mathcal{K} we use must have special properties, such as having a vanishing Ricci tensor. These spaces are known in algebraic geometry under the name of Calabi–Yau manifolds.

While all string theories are equivalent, the situation becomes radically more complicated as soon as compactification is considered. There are many possible choices for these spaces, and a few discriminating criteria. Furthermore, the choice of a Calabi–Yau space leads to precise predictions for the matter content (particle spectrum and possible interactions) and their dynamics.

13.2.5.3 Moduli

Superstring effective theories contain many massless scalar excitation states called ‘moduli’ as they parameterize the space of inequivalent vacua. This space is called the moduli space (some authors also count fermionic states among the moduli, and thus include, for instance, the gravitino, or the supersymmetric partners of the moduli). These states can acquire masses at the symmetry-breaking energy scale, i.e. of the electroweak scale order. Their supersymmetric partners, as well as that of the dilaton (which can itself be classified in the moduli category), can induce the same cosmological problem as the gravitinos.

These moduli are of interest in cosmology because, since string theory contains many of them, regardless of the compactification chosen, they offer many reasonable light-field candidates, for instance, for inflation or for baryogenesis.

However, moduli are in fact problematic for cosmology, and this has come to be known as the Polonyi problem. If the moduli are produced in the primordial Universe, they will behave as matter and decay very late, hence risking dominating the Universe until it is too late for nucleosynthesis to happen in the regime required to satisfy observational constraints.

13.2.6 Origin of three space-like dimensions

One cosmological application of string theory is given by the Brandenberger and Vafa mechanism [10], providing a possible explanation of why our expanding Universe only has three space-like dimensions, while the underlying string theory has nine of them. This mechanism is at the origin of what is known as brane gas cosmology.

13.2.6.1 T-duality

When we consider a set of superstrings described as a gas, unlike what happens for point-like particles, the physics is very different if we consider objects in a box whose characteristic size R is sent to infinity, or if the gas is originally in an infinite space. Both these limits are equivalent for point-like particles. This is due to the fact that particles can only have local interactions, and the global structure of space can have no influence on such a gas. For strings, on the other hand, the situation is different since they are extended objects; as a consequence, if one dimension is compact, they can wind around it, as represented in Fig. 13.5.

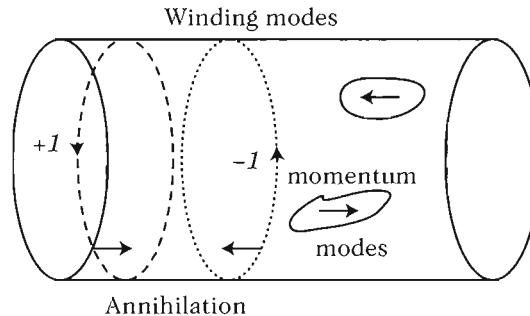


Fig. 13.5 If one dimension is compact, i.e. if it is a circle, two kinds of modes appear. The usual propagation modes, so-called momentum modes, and those for which the string winds around the compact dimension. Winding modes in one way and in the other can annihilate each other.

The propagation momentum modes are excitation states in the compact dimension, and their energy is thus of the form $E_n = n\alpha_m/R$, with $n \in \mathbb{N}$ and α_m a constant depending on the topology of the compact dimension. Indeed, a purely winding mode is more energetic if the dimension is large. As a consequence, it has an energy $E_m = m\alpha_w R$, with here again $m \in \mathbb{N}$ and α_w another constant.

So, the excitation spectrum is invariant under replacing $R \leftrightarrow R^{-1}$ together with $n\alpha_m \leftrightarrow m\alpha_w$. Actually, it is the entire theory that is invariant, and this is simply an illustration of T-duality. More surprisingly, this means that the theory has no way to distinguish between large and small scales.

13.2.6.2 Superstring thermodynamics

String gas thermodynamics gives different results depending on the number of non-compact dimensions. For instance, for at most three non-compact dimensions, there exists a so-called Hagedorn temperature,² at which the total energy diverges. There is no such problem if one considers a compact box whose size is then sent to infinity, because in this case it becomes necessary to include the winding modes. Their energy, which increases as the compactification scale increases, then contributes in such a way as to prevent the total energy to diverge.

Clearly, close to the Hagedorn temperature, the energy fluctuations become very important, and we should resort to a microcanonical description. We then find that, as long as there are some non-compact dimensions, beyond a limiting energy, the specific heats become negative. This implies that string thermodynamics can only be made meaningful in compact spaces. Furthermore, if all the space dimensions are compact, the temperature cannot exceed the Hagedorn temperature: there exists a maximal temperature in this description.

²The temperature of a string gas is defined in the same way as its equivalent for a set of particles, via the coefficient that appears in the distribution function of the string states accessible to the system. It also gives a measure of the characteristic energy of the gas components.

13.2.6.3 Temperature and dimensions

The idea of Brandenberger and Vafa [10] was to note that the space-like dimensions in which our Universe evolves had no reason, to start with, to be different from the others. They hence postulated that the initial conditions on the 9 spatial dimensions should have been the same for all of them, i.e. they were all compact and very small, with a characteristic size of the order of the Planck length. They then studied if some mechanism could exist according to which only three dimensions could start expanding. In this way, the dynamics would naturally lead to cosmology as we know it, namely a universe that appears non-compact, at least on sizes smaller than or comparable to its larger observable distance scales (see Section 12.4.4.6), while explaining why the extra dimensions could not have undergone expansion, thereby remaining compact, and small.

Let us thus consider a universe initially having the topology of a nine-dimensional cubic torus³ of radius R , which will play the role of the scale factor. The expansion can be seen as a relation between this scale factor and the temperature, $T = T(R)$. It is, in fact, not necessary to know the gravitational dynamics to obtain this curve: the adiabaticity condition is sufficient.

First, we know that there is a maximal temperature, and so the curve $T(R)$ does not diverge, contrary to what happens with point-like particles. Moreover, due to the T-duality, we expect the curve to be symmetric under the exchange $\ln R \leftrightarrow -\ln R$. This is exactly what is found in practice: for large R , the temperature decreases as $T \propto R^{-1}$ as we can expect since the momentum modes dominate, whereas for small R , the temperature increases linearly with R because of the winding modes. For intermediate scales, when the temperature is of the order of the Hagedorn temperature, it remains on a plateau.

13.2.6.4 String collisions and four dimensions

Winding modes contribute negatively to the pressure since their energy increases with increasing R . In other words, the winding modes tend to slow down the expansion, or even try to prevent it. Moreover, thermal equilibrium implies that the number of winding modes decreases with the size of the box. So one can only continue the expansion if there is a thermal state that disfavours the winding modes when R increases.

If, for any reason, thermal equilibrium can no longer be maintained for the winding modes, then an important number of them will survive that will slow down and eventually stop the expansion.

To remain in thermal equilibrium, the winding-mode distribution must be able to annihilate itself as in Fig. 13.5. The modes and antimodes should thus be able to approach each other within a distance of the order of the Planck scale. Let us see under which conditions this is possible.

When a string moves, it spans a 2-dimensional space-time surface, its worldsheet. Two given strings in arbitrary states of motion therefore span two worldsheets that have no chance whatsoever to meet in a 9-dimensional space. Therefore, an expanding

³This kind of compactification is not possible in practice in realistic superstring models as they lead, for instance, to real representations for the fermions. Its aim is only to illustrate the idea.

9-dimensional space will eventually come to a halt. With this annihilation mechanism, the maximal dimensionality of the space in which we expect string collisions is given by the two surfaces covered by the strings. One can easily convince oneself that in a 3-dimensional space such as ours, winding modes will generically be able to meet each other, and hence to annihilate. In fact, three is the largest possible number of dimensions that can start expanding because of this mechanism. In other words, according to this analysis, our Universe has three dimensions because the fundamental objects are strings, with one spatial dimension: assuming one-dimensional strings as the fundamental objects implies not only that the number of dimensions is fixed for mathematical consistency reasons at the quantum level, it also naturally explains why our Universe is seen as a 4-dimensional space-time, merely as the result of the early cosmological evolution.

13.2.6.5 Brane gas

The ideas developed above to explain the number of dimensions of our space-time can also be discussed in the context of \mathcal{M} -theory in 11 dimensions; this is the brane gas scenario [11]. One can show that the fundamental degrees of freedom are then the graviton, 2-branes and 5-branes. Compactifying the eleventh dimension on a circle, we then find that the possible states are the 0-branes, the strings (1-brane), as well as 2-, 4-, 5-, 6-, and 8-branes. We now assume that the initial state has, by symmetry, as many winding as antiwinding modes, and that the Universe has a toroidal topology where all the dimensions behave in a similar way. When the scale factor a starts growing, since the energy of a p -brane is proportional to a^p , the larger the value of p , the more important are the effects of these p -branes. So it is the large dimensionality p -branes that start interacting first. Yet, we can show that p -branes can only interact and, as a consequence, annihilate in a space of dimension $2p + 1$ at most. In other words, with $D = 10$ and thus 9 space-like dimensions, all the p -branes with $p \leq 4$ will quickly annihilate. Only the strings remain, and the Brandenberger–Vafa mechanism applies straightforwardly.

13.3 The Universe as a ‘brane’

Among the many consequences of string theories, we find the existence of extra dimensions, usually assumed to be compact. We also find new objects, called D p -branes. These branes have opened a rich phenomenology for primordial cosmology [12–16].

13.3.1 Motivations

13.3.1.1 Hořava–Witten realization

In 1996, Hořava and Witten [17] realized that \mathcal{M} -theory, once compactified on the space $\mathbb{R}^{10} \times S^1/Z_2 = \mathbb{R}^{10} \times [0, 1]$ leads to the strong coupling limit of heterotic string theory $E_8 \times E_8$, provided the gauge fields are in a 10-dimensional vector multiplet that only propagates on the space-time boundaries. Gravity, on the other hand, propagates in all the dimensions. The picture this leads to is that of supergravity in an 11-dimensional space-time (the bulk) with the eleventh dimension bounded by two 10-dimensional boundary branes where the gauge fields of the group E_8 evolve. One

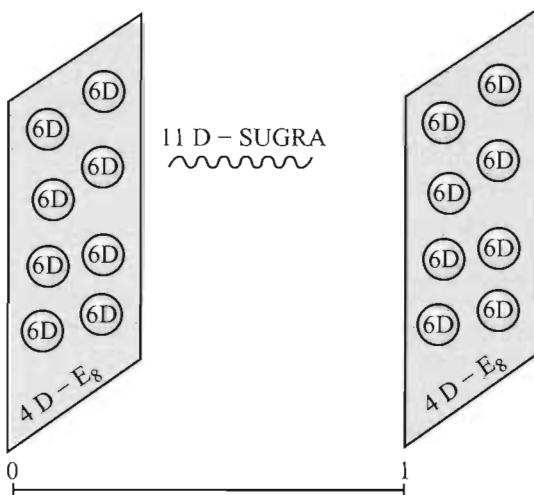


Fig. 13.6 Hořava-Witten model: the 11-dimensional space-time of \mathcal{M} -theory is bounded by two 10-dimensional branes. Each of these boundary branes is in turn decomposed into a large 4-dimensional brane and 6-dimensional compact spaces at every point (the circles represent the Calabi-Yau manifolds at each point of the brane). Supergravity exists here in all 11 dimensions.

further needs to compactify the six additional dimensions in the usual way on a Calabi-Yau manifold to obtain the picture of Fig. 13.6. The eleventh dimension, or the fifth once the 6 internal dimensions are compactified, is not necessarily strongly limited in size since all the known particle fields are only supposed to exist on the branes, and only gravity exists in all the dimensions. As a consequence, only gravity is constrained. These constraints indicate that the extra dimensions should be smaller than about a millimeter (see Section 1.6).

Although this picture leads to a cosmological framework that is conceptually very far from that of the standard model based on 4-dimensional general relativity, it can, however, fortunately be made to contain this standard model (otherwise it would already be ruled out). Furthermore, it can be generalized by assuming that our Universe is a 3-brane embedded in a space-time of arbitrary dimension (Fig. 13.7).

13.3.1.2 The hierarchy problem

Another key theoretical problem, that can in no way be addressed in the framework of either standard cosmology or particle physics, is the hierarchy problem, i.e. why is the gravitational scale $M_p = G_N^{-1/2}$ so far away from that which is characteristic of the electroweak interactions.

In 1998, Arkani-Hamed et al. [18] suggested a mechanism thanks to which this problem might not be beyond the reach of a scientific explanation. Their solution, as it turns out, is also based on the existence of extra dimensions, in their case at least one large one.

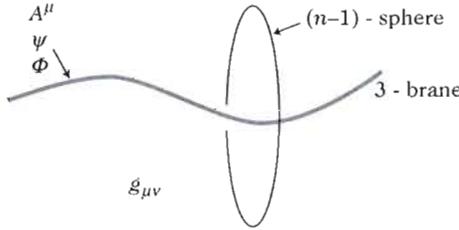


Fig. 13.7 A brane can be seen as a special configuration (a topological defect, for instance, or the set of singular points on a compactification manifold) in three dimensions (if it must describe our space, but there are other possibilities in string theory) existing in a higher-dimensional space-time. While the fields describing matter, the gauge fields A^μ , spinor fields ψ and scalar fields Φ evolve as being strictly confined to the brane, each point of it is surrounded by a space, called the ‘bulk’, in which we only find the effects of gravity $g_{\mu\nu}$.

The starting idea is to assume that the fundamental energy scale is that of the electroweak interactions m_{ew} , and that there are n compact dimensions, of radius R . The Planck mass in the $4+n$ dimensions is therefore chosen to be $M_P^{(4+n)} \sim m_{ew}$. The Newtonian potential $V_<(r)$ for two test masses m_1 and m_2 separated by $r \ll R$ is then

$$V_<(r) \simeq \frac{m_1 m_2}{[M_P^{(4+n)}]^{n+2}} \frac{1}{r^{n+1}}.$$

This expression can be compared to the potential $V_>(r)$ for two masses separated by a distance $r \gg R$, i.e. sufficiently larger than the radius of the extra dimensions that the masses do not feel the effects of these dimensions. It is

$$V_>(r) \simeq \frac{m_1 m_2}{M_P^2} \frac{1}{r} \simeq \frac{m_1 m_2}{R^n [M_P^{(4+n)}]^{n+2}} \frac{1}{r},$$

which is another equivalent way to obtain the relation between the gravitational constants in 4 and $4+n$ dimensions [see below (13.12)]. We then find $M_P^2 = [M_P^{(4+n)}]^{n+2} R^n$, or rather

$$R \simeq 10^{-17+30/n} \text{ cm} \times \left(\frac{1 \text{ TeV}}{m_{ew}} \right)^{1+2/n},$$

for the characteristic radius of the extra dimensions. Here, we assume that the hierarchy problem is solved by considering an energy scale of the order of a TeV, i.e. not very far from the characteristic scales of the standard model of particle physics.

Recalling the experimental constraint $R \leq 1 \text{ mm}$, the result above implies that there should be at least two extra dimensions. As it turns out, the case $n = 2$ is the most interesting one as it leads to predictions on scales of the order of the millimeter, which is within the limits of current experiments: if this mechanism is indeed responsible for the large difference between the Planck and electroweak scales and if it was realized using two large extra dimensions, then we should soon be able to detect some experimental consequences and effectively prove that we live in more than 3 spatial dimensions.

13.3.2 Induced Einstein equations

Let us consider an arbitrary brane, i.e. a p -dimensional space embedded in an n -dimensional space ($n > p$), with coordinates x^A and metric \bar{g}_{AB} , $A, B = 0, \dots, n - 1$. The junction conditions discussed in Section 8.4.5 are easily generalized. To do so, we set $x^A = X^A(x^\mu)$ the brane coordinates, with internal coordinates x^μ , with $\mu = 0, \dots, p - 1$.

13.3.2.1 Fundamental forms

With the induced metric, $g_{\mu\nu} = \bar{g}_{AB}X_{,\mu}^AX_{,\nu}^B$, one defines the first fundamental form $\eta^{AB} = g^{\mu\nu}X_{,\mu}^AX_{,\nu}^B$, as well as the second fundamental form

$$K_{AB}^C \equiv \eta_B^D\eta_A^E\nabla_E\eta_D^C. \quad (13.37)$$

This tensor is tangent to the brane in its two first indices and orthogonal in the third one, namely

$$\perp_D^K K_{AB}^C = \perp_D^K K_{BA}^C = \eta_D^C K_{BA}^D = 0, \quad \text{with} \quad \perp_{AB} = \bar{g}_{AB} - \eta_{AB}.$$

In order for the brane to effectively represent a p -volume in the entire space, the second fundamental tensor must be symmetric in its first two indices,

$$K_{BA}^C = K_{AB}^C.$$

This relation, called the Weingarten identity, provides a geometrical equation of motion for the brane.

From now on, we shall restrict attention to the case of the hypersurface already discussed in Section 8.4.5, specializing to the 3-surface embedded in a five-dimensional space-time. In this case with only one codimension, one can define uniquely a space-like unit vector n^A , normal to the brane ($\bar{g}_{AB}n^An^B = 1$), with which one builds the projection tensor orthogonal to the surface as $\perp_B^A = n^An_B$. One can then write the extrinsic curvature as a function of the second fundamental form, namely

$$K_{AB}^C = K_{AB}n^C, \quad \text{with} \quad K_{AB} = -\eta_B^D\eta_A^E\nabla_E n_D = -\nabla_A n_B.$$

This is equivalent to (8.160), where the last equality is only valid if n_A is continuous in the bulk along the geodesics, i.e. $n^A\nabla_A n_B = 0$. The case of a single extra dimension is the only case for which simple junction conditions can be used in an unambiguous way, and we will thus restrict ourselves to this case from now on.

The Weingarten identity for the second fundamental form of the brane implies that the extrinsic curvature K_{AB} must be symmetric. This arises from the fact that the extrinsic curvature can also be defined via the Lie derivative of the metric along the direction orthogonal to the brane [19], i.e. $K_{AB} = \frac{1}{2}\mathcal{L}_n\bar{g}_{AB}$.

Finally, let us mention the Codazzi equation, which is

$$\nabla_A K_B^A - \nabla_B K = \bar{R}_{AC}\eta_B^A n^C, \quad (13.38)$$

determining the variation rate of the extrinsic curvature along the lines of constant y .

13.3.2.2 Intrinsic curvature and junction conditions

The intrinsic curvature, $R_{\mu\nu}$, can be related to the embedding one, \bar{R}_{AB} , through the Gauss relation

$$\mathcal{R}_{AB} = \frac{n-3}{n-2}\bar{R}_{CD}\eta_A^C\eta_B^D + \frac{1}{n-2}\bar{R}_{CD}\eta^{CD}\eta_{AB} - \frac{1}{n-1}\bar{R}\eta_{AB} + KK_{AB} - K_A^CK_{BC} + \bar{\mathcal{W}}_{AB}, \quad (13.39)$$

where $\mathcal{R}^{AB} = R^{\mu\nu}X_{,\mu}^AX_{,\nu}^B$, and $\bar{\mathcal{W}}_{AB}$ is the orthogonal projection of the embedding space Weyl tensor, $\bar{\mathcal{W}}_{AB} = \mathcal{C}_{ACBD}\perp^{CD}$.

Although we must demand the metric to be continuous in the entire space, there is no such necessary constraint for its derivatives that can change value across the brane. Modelling the latter by a localized energy-momentum tensor

$$T^{AB}\Big|_{\text{brane}} = \int d^p x \sqrt{-g} \bar{T}^{AB} \delta^{(n)}[x^A - X^A(x^\mu)], \quad (13.40)$$

just as in the cosmic-string case of Chapter 10, we set

$$\langle K_{AB} \rangle \equiv \frac{1}{2}(K_{AB}^+ + K_{AB}^-) \quad \text{and} \quad [K_{AB}] \equiv K_{AB}^+ - K_{AB}^-,$$

where the + and – indices indicate that the tensor is evaluated on one side and the other of the hypersurface.

The junction condition (8.162) then becomes

$$[K_{AB}] = -\kappa_n \left(\bar{T}_{AB} + \frac{1}{1-p} \bar{T} \eta_{AB} \right), \quad (13.41)$$

where $\bar{T} = \bar{g}^{AB}\bar{T}_{AB}$ and κ_n is given by (13.2). We obtain the brane internal dynamics from this equation: in a cosmological context for which one assumes a brane metric of the FLRW form, this leads to a modified Friedmann equation.

In the case where the brane is subject to an exterior force $f^A = fn^A$, a generalization of the second Newton's law leads to

$$T^{AB}\langle K_{AB} \rangle = f, \quad (13.42)$$

a relation that gives the dynamics of the brane in the entire space. This equation can be obtained from the Gauss equation (13.39) and the continuity of \mathcal{R}_{AB} across the brane.

13.3.2.3 Z_2 -symmetry

In many cases, either for the sake of simplicity or because the Hořava–Witten model possesses such a symmetry, we consider the brane to be localized at $y = 0$ and we are interested in a symmetric configuration, i.e. one invariant under the transformation $y \rightarrow -y$. Because of this symmetry, there can be no exterior force, and therefore we must set $f = 0$. As a result, the extrinsic curvature tensor K_{AB} changes sign from one side of the brane to the other, so that $[K_{AB}] = 2K_{AB}(0)$, and the junction condition

(13.41) then implies that the brane extrinsic curvature is completely determined by its matter content.

Let us consider explicitly the case of a 4-dimensional brane ($p = 4$ in the discussion above) embedded in a 5-dimensional manifold ($n = 5$ then). Using (13.39) and the 5-dimensional Einstein equations

$$\bar{G}_{AB} + \Lambda_5 \bar{g}_{AB} = \kappa_5 T_{AB}, \quad (13.43)$$

where to remain general we have kept a cosmological constant Λ_5 , we can obtain the induced four-dimensional Einstein equations by decomposing the metric as

$$ds^2 = g_{\mu\nu}(x^\alpha, y) dx^\mu dx^\nu + dy^2,$$

which leads to

$$G_{\mu\nu} + \lambda g_{\mu\nu} = \kappa_4 T_{\mu\nu} + 6 \frac{\kappa_4}{\lambda} S_{\mu\nu} - \mathcal{E}_{\mu\nu} + 4 \frac{\kappa_4}{\lambda} \mathcal{F}_{\mu\nu}. \quad (13.44)$$

To obtain this relation, we have considered that the total energy-momentum tensor of the brane itself contains a 4-dimensional cosmological constant part, i.e. $T_{\mu\nu}^{\text{brane}} = T_{\mu\nu} - \lambda g_{\mu\nu}$, and we then have $\kappa_4 = \frac{1}{6}\lambda\kappa_5$ as well as $\Lambda_4 = \frac{1}{2}(\Lambda_5 + \kappa_4\lambda)$. It is interesting to note that in the absence of such a four-dimensional cosmological constant, we do not recover Einstein equations. To obtain (13.44), we should use the junction condition (13.41) allowing us to express the extrinsic curvature as a function of the energy-momentum tensor.

The corrections to Einstein equations come first from the extrinsic curvature

$$S_{\mu\nu} = \frac{T}{12} T_{\mu\nu} - \frac{1}{4} T_{\mu\alpha} T_\nu^\alpha + \frac{1}{24} g_{\mu\nu} (3T_{\alpha\beta} T^{\alpha\beta} - T^2), \quad (13.45)$$

which is quadratic in the source. A second contribution arises from the Weyl tensor $\mathcal{E}_{\mu\nu} = \bar{W}_{AB} \eta_\mu^A \eta_\nu^B$. Finally, a term coming from the bulk

$$\mathcal{F}_{\mu\nu} = T_{AB} \eta_\mu^A \eta_\nu^B + \left(T_{AB} n^A n^B - \frac{1}{4} T \right) g_{\mu\nu}, \quad (13.46)$$

in which T represents an energy-momentum source that exists only in the bulk. In most cosmological applications, this term is assumed to be absent, so we can set $T_{AB} = 0$.

13.3.2.4 Conservation of $T_{\mu\nu}$

The Codazzi equation (13.38), together with the junction condition (13.41), implies that

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

so that the energy-momentum tensor remains conserved on the brane. As a consequence, there is no energy and momentum exchange between the brane and the bulk other than purely gravitational.

13.3.3 The Randall–Sundrum model

13.3.3.1 An infinite dimension

Let us assume that the bulk has a negative cosmological constant, $\Lambda_5 < 0$. In this case, the bulk global structure is that of a 5-dimensional anti-de Sitter space. For a brane with no matter, one then obtains the solution

$$ds^2 = e^{-2|y|/\ell} \eta_{\mu\nu} dx^\mu dx^\nu + dy^2, \quad \text{with } \ell = \sqrt{-\frac{6}{\Lambda_5}}, \quad (13.47)$$

which is called a ‘warped’ solution. Such a solution preserves the Lorentz invariance and the four-dimensional Minkowski space structure of the brane. This solution is also comparable to that of considering a compact extra dimension: because of the anti-de Sitter structure, and despite the fact that the extra dimension y varies in \mathbb{R} , the bulk is of finite volume. Indeed, integrating the exponential factor of the metric along the fifth dimension, we find that

$$\int d^5x \sqrt{-\bar{g}} = \ell \int d^4x \sqrt{-g},$$

and similarly for the curvature, so that the Planck masses in four and five dimensions are now related by $[M_p^{(5)}]^3 = 2 [M_p^{(4)}]^2 / \ell$.

On the brane, the metric of the solution (13.47) is a Minkowski metric. This means that the four-dimensional cosmological constant must effectively vanish. We should therefore have $\Lambda_5 = -\kappa_4 \lambda$, i.e. the brane tension λ should be fixed by the bulk curvature radius as

$$\lambda = \frac{3M_p}{4\pi\ell^2}, \quad (13.48)$$

which represents a parameter fine tuning that is hardly justifiable at this level.

13.3.3.2 Two branes

Another way to model space compactification is to use a second brane, located at a distance $y = L$, and assuming again the Z_2 -symmetry. Note that in this case, this symmetry now leads to invariance under both the transformations $y \rightarrow -y$ and $y+L \rightarrow L-y$.

For such a solution to be meaningful and satisfy the symmetry requirements, both branes should have equal and opposite tensions, namely $\pm\lambda$, as defined by (13.48). Moreover, the effective Planck scale should now be replaced by

$$[M_p^{(4)}]^2 = [M_p^{(5)}]^3 \ell \left(1 - e^{-2L/\ell}\right). \quad (13.49)$$

In this model, where one assumes that our Universe is the brane located at $y = L$, we recover at low energy that gravity is described by a scalar-tensor theory. To recover 4-dimensional general relativity, one must find a way to stabilize the distance between the branes. This distance is seen from the brane point of view as a scalar field, called the radion. A stabilized radion is necessary for the model to make predictions compatible with observations.

13.3.3.3 Gravitational attraction

From now on, we only consider the single-brane case, which is equivalent to taking the limit $L \rightarrow \infty$ in the two-brane case.

If there is matter on the brane, we can look for a solution of the form

$$ds^2 = e^{-2|y|/\ell} \eta_{\mu\nu} dx^\mu dx^\nu + dy^2 + h_{AB} dx^A dx^B, \quad (13.50)$$

where h_{AB} is a perturbation.

In a way similar to that for cosmological perturbations, the equations of motion obtained for h_{AB} are invariant under gauge transformations of h_{AB} . We should therefore choose a gauge before entering the details of the calculations. It turns out to be convenient to assume the so-called ‘transverse and traceless’ (TT) gauge, defined by

$$\partial_\mu h^{\mu\nu} = h^\mu_\mu = 0,$$

which is, however, not quite sufficient. There remain some degrees of freedom to be fixed, and this can be done by setting

$$h_{\mu y} = h_{yy} = 0,$$

called the Gaussian normal gauge.

With this gauge choice, Einstein’s equations read

$$\left(e^{-2|y|/\ell} \square + \partial_y^2 - \frac{4}{\ell^2} \right) h_{\mu\nu} = 0. \quad (13.51)$$

These gauge choices completely fix the brane position at $y = \mathcal{F}^4(x^\alpha)$, i.e. a yet-unknown function of the brane internal coordinates. But we can also choose, as this merely amounts to redefining the extra coordinate, to use the junction conditions (13.41) to impose that the brane is localized at $y = 0$, which often simplifies the computation. To do so, we only keep the Gaussian normal gauge conditions and we find

$$\left(\partial_y + \frac{2}{\ell} \right) \tilde{h}_{\mu\nu} = -\kappa_5 \left(T_{\mu\nu} - \frac{e^{-2|y|/\ell}}{3} \eta_{\mu\nu} T \right), \quad (13.52)$$

where the perturbation is now denoted by \tilde{h} .

One can switch from one gauge to the other by making the change of coordinates generated by a vector v^A of the form

$$v^4 = \mathcal{F}^4(x^\alpha), \quad v^\mu = \mathcal{F}^\mu(x^\alpha) - \frac{\ell}{2} e^{-2|y|/\ell} \partial^\mu \mathcal{F}^4(x^\alpha),$$

where \mathcal{F}^μ are functions of the brane internal coordinates. The relation between the perturbation variables in the two gauges is then given by

$$h_{\mu\nu} = \tilde{h}_{\mu\nu} - \ell \partial_\mu \partial_\nu \mathcal{F}^4 - e^{-2|y|/\ell} \left(\frac{2}{\ell} \eta_{\mu\nu} \mathcal{F}^4 - \partial_{(\mu} \mathcal{F}_{\nu)} \right),$$

thanks to which we find the junction condition for h

$$\left(\partial_y + \frac{2}{\ell}\right) h_{\mu\nu} = -\kappa_5 \left(T_{\mu\nu} - \frac{e^{-2|y|/\ell}}{3} \eta_{\mu\nu} T\right) - 2\partial_\mu \partial_\nu \mathcal{F}^4. \quad (13.53)$$

This condition can now be combined with (13.51) to give

$$\left(e^{2|y|/\ell} \square + \partial_y^2 - \frac{4}{\ell^2} + \frac{4}{\ell} \delta(y)\right) h_{\mu\nu} = -2\kappa_5 S_{\mu\nu} \delta(t), \quad (13.54)$$

where

$$S_{\mu\nu} = \left(T_{\mu\nu} - \frac{e^{-2|y|/\ell}}{3} \eta_{\mu\nu} T\right) + \frac{2}{\kappa_5} \partial_\mu \partial_\nu \mathcal{F}^4.$$

The general solutions of this relation can be expressed in terms of retarded Green functions evaluated on the brane. These Green functions can be explicitly expanded into massive and massless modes, the latter playing the role of the graviton, whereas the former allow for corrections due to the extra dimension. We find

$$G(\mathbf{x}, y, \mathbf{x}', y') = \int \frac{d^4 k}{(2\pi)^4} e^{ik \cdot (\mathbf{x}-\mathbf{x}')} \left[\frac{1}{\ell} \frac{e^{-2(|y|+|y'|)/\ell}}{(k_0 + i\epsilon)^2 - \mathbf{k}^2} + \frac{u_m(y) u_m(y')}{(k_0 + i\epsilon)^2 - \mathbf{k}^2 - m^2} \right], \quad (13.55)$$

where the modes u_m are given by properly normalized Bessel functions, namely

$$u_m(y) = \sqrt{\frac{m\ell}{2}} \frac{J_1(m\ell) Y_2(e^{|y|/\ell} m\ell) - Y_1(m\ell) J_2(e^{|y|/\ell} m\ell)}{\sqrt{[J_1(m\ell)]^2 + [Y_1(m\ell)]^2}}.$$

The solution for h is then

$$h_{\mu\nu} = -2\kappa_5 \int G(\mathbf{x}, 0, \mathbf{x}', 0) S_{\mu\nu}(\mathbf{x}', 0) d^4 x',$$

so that we recover the TT gauge conditions provided $\square \mathcal{F}^4 = \frac{1}{6} \kappa_5 T$.

One can choose the functions \mathcal{F}^μ in such a way as to also have

$$\tilde{h}_{\mu\nu} = -2\kappa_5 \int G(\mathbf{x}, 0, \mathbf{x}', 0) \left(T_{\mu\nu} - \frac{e^{-2|y|/\ell}}{3} \eta_{\mu\nu} T\right) d^4 x' + \frac{2}{\ell} e^{-2|y|/\ell} \eta_{\mu\nu} \mathcal{F}^4,$$

and therefore, one exactly recovers the linearized Einstein equations for these perturbations provided we truncate the Green function to keep only the massless modes.

To study the proper gravitational interaction between the particles on the brane, we take an energy-momentum tensor of the form $T_{\mu\nu} = \rho u_\mu u_\nu$, where the u_μ are four-velocity vectors. Assuming spherical symmetry for the source, we find the potential

$$\frac{1}{2} \tilde{h}_{00} = \frac{G_N M}{r} \left(1 + \frac{2\ell^2}{3r^2}\right), \quad (13.56)$$

with $r = |\mathbf{x} - \mathbf{x}'|$ and we have set $M \equiv \int \rho d^3 x$ the source mass. The modifications to Newton's law are manifest on scales smaller than or comparable to ℓ .

13.3.4 Cosmological phenomenology

13.3.4.1 Choice of coordinates

The bulk solutions we will be interested in, in a cosmological context, are those allowing to recover a Friedmann–Lemaître metric on the brane. It was shown that the metric can be set in the form

$$ds^2 = -f(R)dt^2 + \frac{dR^2}{f(R)} + R^2 \left(\frac{dr^2}{1-Kr^2} + r^2 d\Omega^2 \right), \quad (13.57)$$

where the function $f(R)$ is given by $f(R) = K + R^2/\ell^2$, the constant $K = 0, \pm 1$ being the three-dimensional spatial curvature. As the brane moves through the bulk, the solution of Einstein equations leads to $R = a(T)$, i.e. a given function of time. The quantity $a(T)$ is eventually interpreted as the brane scale factor, so that the expansion appears to be induced by the motion of the brane through the bulk.

Changing to Gaussian normal coordinates, we have

$$ds^2 = -N^2(t, y)dt^2 + A^2(t, y) \left(\frac{dr^2}{1-Kr^2} + r^2 d\Omega^2 \right) + dy^2, \quad (13.58)$$

and the scale factor is now identified to $a(t) = A(t, 0)$. As long as t is the brane proper time, we can also set $N(t, 0) = 1$, and we then find

$$N = \frac{\dot{A}(t, y)}{\dot{a}(t)} \quad \text{as well as} \quad A = a(t) \left\{ \cosh \left(\frac{y}{\ell} \right) - \left[1 + \frac{\rho(t)}{\lambda} \right] \sinh \left(\frac{|y|}{\ell} \right) \right\},$$

so that the induced Einstein equation now reads

$$H^2 = \frac{\kappa}{3} \rho \left(1 + \frac{\rho}{\lambda} \right) + \frac{\Lambda}{3} - \frac{K}{a^2}, \quad (13.59)$$

and hence a modified Friedmann equation [20].

13.3.4.2 Fluid solutions

To understand quantitatively in what respect (13.59) differs from the usual Friedmann equation, let us compute its solutions when the source term is a perfect fluid. Assuming that the equation of state is w , and because the conservation equation is unchanged on the brane, one finds that the relation for $\rho(a)$ is not modified. We then have $\rho \propto a^{-3(1+w)}$. Setting for simplicity $\Lambda = 0$, and considering Euclidean spatial sections, i.e. $K = 0$, we get the exact solution in the form

$$a(t) \propto [t(t + t_\lambda)]^{1/3(1+w)}, \quad \text{with} \quad t_\lambda = \frac{M_p}{\sqrt{3\pi\lambda(1+w)}},$$

which involves a characteristic time t_λ given by the microphysics as well as the matter content of the brane. The corrections induced are thus significant only for $t \ll t_\lambda$, and so the solution is naturally driven to the standard one on large timescales.

Note that for $w = -1$, we recover that ρ should be constant in time, so that the de Sitter solution is not modified. The only difference merely stems from the actual value of the Hubble constant that differs from its 4-dimensional counterpart.

13.3.4.3 Brane inflation

In order for brane cosmology to reproduce the standard cosmological model, an inflationary phase must have occurred on the brane. To realize such a phase, many methods have been proposed, using brane or bulk scalar fields, or interactions between both branes. They are detailed in Ref. [21], and we provide a glimpse of these ideas below.

We assume that only gravity propagates in the bulk, and hence consider a scalar field ϕ confined to the brane, whose dynamics stems from a potential $V(\phi)$. With the idea of achieving an inflationary phase, we first note that the high-energy corrections to the Friedmann equation actually increase the Hubble damping. This renders slow-roll inflation more probable: inflation can in fact take place even for potentials that would not have led to it in usual cosmology.

Indeed, the condition for the Universe to accelerate, i.e. $\ddot{a} > 0$, translates for the equation of state into

$$w < -\frac{1}{3} \frac{\lambda + 2\rho}{\lambda + \rho},$$

which is a priori more constraining than the usual constraint. In the slow-roll approximation, we find

$$H^2 \simeq \frac{\kappa}{3} V \left(1 + \frac{V}{\lambda} \right),$$

and

$$\dot{\phi} = -\frac{V_{,\phi}}{3H}.$$

We clearly see that the expansion rate H is slightly larger than it would be in the absence of the correction term. Therefore, the scalar field motion is more damped, and the slow-roll regime is easier to reach. This translates into effectively smaller slow-roll parameters than in the four-dimensional case with the same potential.

13.3.5 Possible extensions

Different paths have been followed to extend the previous considerations on brane physics. They all amount to considering the minimal theory presented so far, adding new terms to include the possible effects.

13.3.5.1 On the reflection symmetry

The first possibility one can think of is to break the Z_2 -symmetry across the brane. In this case, the energy-momentum tensor no longer has any reason to be identical on each side of the brane, leading to a force term in (13.42) expressible as

$$f = -[\mathcal{T}^{AB}] n_A n_B.$$

In this category of models, one recovers the fact that linearized general relativity is still a good approximation.

An example of implementation of this breaking of the Z_2 -symmetry is provided by assuming that the brane is charged and that it couples to a 4-form, $A_{\alpha\beta\gamma\delta}$ defined in the bulk. The equation of motion for such a form, namely $\nabla_{[\mu} A_{\alpha\beta\gamma\delta]} = F \epsilon_{\mu\alpha\beta\gamma\delta}$ where F is a pseudo-scalar, implies that its exterior derivative is constant. The jump in F

across the brane is proportional to the brane charge so that the effective cosmological constant is different on each side of the brane [22].

13.3.5.2 Gauss–Bonnet term

In five or more dimensions, one can add a term that, although quadratic in the curvature, does not change the order of the equations of motion: a Gauss–Bonnet term,

$$\mathcal{L}_{\text{GB}} \propto \bar{R}^2 - 4\bar{R}_{AB}\bar{R}^{AB} + \bar{R}_{ABCD}\bar{R}^{ABCD}. \quad (13.60)$$

Note that in four dimensions, this term is a total derivative, and cannot change the dynamics. On the other hand, in more than four dimensions, the dynamics is explicitly modified by (13.60), so that we expect new effects. For instance, the coherence relation (8.218) between the scalar and tensor spectra gets modified [23]. It is also believed that such a term could provide a means of obtaining consistent junction conditions in spaces of codimension higher than one [24].

13.3.5.3 Bulk scalar field

Finally, modifications to the brane model in which one or more scalar fields are free to move in the bulk instead of being confined to the brane have been proposed. Such a modification is rather well motivated and, in fact, could turn out to be necessary for stability reasons. Models of this category lead to a gravitation theory on the brane that is of the scalar-tensor type [25]. They are therefore more constrained.

13.3.6 The Universe as a defect

Having opened the possibility that fields other than the gravitons could propagate in the bulk, one can use them to form topological defects. In the case of one codimension, these must be similar to domain walls. Although the width may not necessarily be negligible, one can, however, imagine a configuration in which the topological defect itself is understood as the brane, the fields induced inside having different properties and behaviours from the bulk ones [26]: for instance, as in the case of a superconducting cosmic string, a symmetry could be broken inside the defect and restored outside (or the other way round), meaning the mass spectrum could change as one goes from inside the defect to the bulk. Let us see this on a simple 5-dimensional model.

13.3.6.1 A simplistic model

The easiest way to identify a brane to a domain wall consists in breaking a Z_2 -symmetry by means of 5-dimensional Higgs field. We then set our action as

$$S = \int \left[\frac{1}{2\kappa_5} (R - 2\Lambda) - \frac{1}{2} g^{AB} \partial_A \Phi \partial_B \Phi - V(\Phi) \right] \sqrt{-g} d^5x, \quad (13.61)$$

demand that the potential be

$$V(\Phi) = \frac{\lambda}{8} (\Phi^2 - \eta^2)^2, \quad (13.62)$$

and choose a warped metric for the infinite extra dimension

$$ds^2 = g_{AB} dx^A dx^B = -dy^2 + e^{-2\sigma(y)} \eta_{\mu\nu} dx^\mu dx^\nu = -dy^2 + g_{\mu\nu} dx^\mu dx^\nu. \quad (13.63)$$

With this metric, Einstein equations for a static configuration are written in the form

$$\frac{3}{\kappa_5} \sigma'' = \Phi'^2 \quad \text{and} \quad 6\sigma'^2 = \frac{\kappa_5}{2} (\Phi'^2 - 2V) - \Lambda, \quad (13.64)$$

while the Klein–Gordon equation becomes

$$\Phi'' - 4\sigma' \Phi' = \frac{dV}{d\Phi}, \quad (13.65)$$

in which a prime denotes a derivative with respect to y , $' \equiv \partial_y$.

13.3.6.2 Boundary conditions

The boundary conditions needed to describe a 4-dimensional domain wall embedded in a 5-dimensional antide Sitter space are the trivial generalizations of that described in Chapter 11. We require that the Higgs field vanishes on a hypersurface, which by convention we choose to be located at $y = 0$. Moreover, we want the symmetry to be broken far from the brane, and as a consequence, we impose $\lim_{y \rightarrow \pm\infty} \Phi = \pm\eta$. The choice of sign is completely arbitrary and corresponds to what we call a *kink*; the opposite choice, i.e. $\lim_{y \rightarrow \pm\infty} \Phi = \mp\eta$, would be an *antikink*, whose physical properties are exactly identical, from the point of view of the brane physics and as long as there is only one configuration of this kind in play.

As for the metric function σ , it must tend asymptotically to an antide Sitter space, implying that σ' must tend towards a constant, that can be determined from (13.64). We find that we should impose $\lim_{y \rightarrow \pm\infty} \sigma' = \pm\sqrt{-\Lambda/6}$.

Using the redefinitions $\varrho \equiv y\sqrt{|\Lambda|}$, $H \equiv \Phi/\eta$ and $S \equiv d\sigma/d\varrho$, one can write the dynamical equations in the form

$$\dot{S} = \frac{\alpha}{3} \dot{H}^2, \quad \ddot{H} - 4S\dot{H} = 4\beta H(H^2 - 1), \quad (13.66)$$

where $\dot{H} = dH/d\varrho$, and where the two dimensionless constants are defined by

$$\alpha \equiv \kappa_5^2 \eta^2, \quad \beta \equiv \frac{\lambda\eta^2}{8|\Lambda|}. \quad (13.67)$$

13.3.6.3 Tuning of the parameters of the solution

Finally, we require that the brane configuration respects the Z_2 -symmetry between y and $-y$. This implies that we need to impose $S(0) = 0$, or $\lim_{\varrho \rightarrow -\infty} S(\varrho) = -1/\sqrt{6}$, which represents a constraint. To satisfy it, there must exist a relationship between the two parameters α and β . It turns out that they must lie on the curve represented in Fig. 13.8, leading to the type of solution also represented in the same figure.

Once the tuning is made (which is approximated by $\alpha^2\beta^2 = 1 + \frac{16}{9}\beta$, as can be seen in figure 13.8), it translates into a relation between the bulk cosmological constant, the brane energy density, and the 5-dimensional Planck mass. In other words, it provides a microscopic fine-tuning interpretation for (13.48).

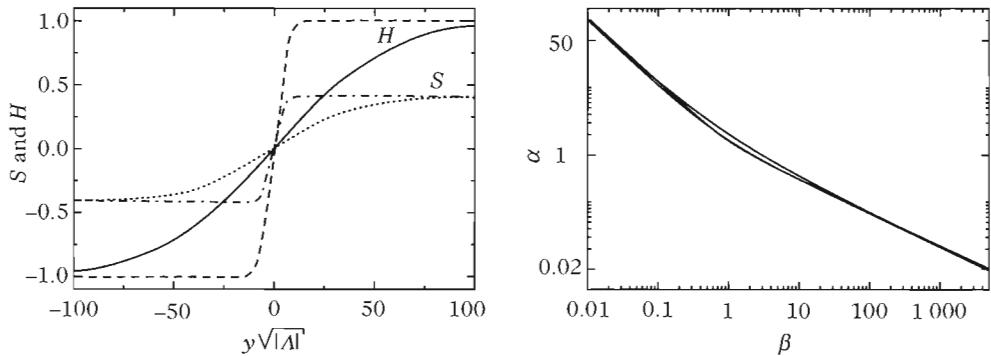


Fig. 13.8 Brane domain-wall solutions and parameter tuning. On the left, the Higgs fields H , as the plain line for $\beta = 0.01$ and dashed for $\beta = 0.1$, and the metric coefficient S , respectively, in dotted and dashed-dotted lines for the same values of β . Once the value of β is fixed, the value of α is fixed so that $S(0) = 0$, leading to the right curve (on which the approximation is represented in the thinner line than the real numerical solution).

13.3.6.4 Extensions

One can generalize the discussion above by considering higher-order topological defects such as hyperstrings or hypermonopoles, of, respectively, codimension 2 and 3. This means that one chooses anti de Sitter spaces with, respectively, 6 or 7 dimensions, where the extra dimensions are still warped. Note that this implies that if we want the defect to be well localized, we should take gauge fields into account in addition to the scalar fields. This leads to models that become extremely complicated very rapidly.

13.3.7 Models of induced gravity

There is yet another category of models with one extra dimension, in which this dimension is infinite and flat, i.e. not warped. Note by the way that it allows for a supersymmetric theory in the bulk, supersymmetry being broken on the brane. We consider models [27] in which, in addition to the five-dimensional gravity action, quantum corrections, usually due to strings, predict an extra term, so that the total action is given by

$$S_{\text{induced}} = \frac{1}{2\kappa_5} \int d^5x \sqrt{-\bar{g}} \bar{R} + \int d^4x \sqrt{-g} \left(\frac{R}{2\kappa_4} + \mathcal{L}_{\text{matter}} \right), \quad (13.68)$$

where the second term is purely four-dimensional, and the metric induced in 4 dimensions is $g_{\mu\nu} = \bar{g}_{\mu\nu}(x, y = 0)$. The matter Lagrangian contained on the brane can be arbitrary, and in particular it can serve as a source for the induced metric.

In order to know how gravity acts on the brane, we consider the total metric to be $\bar{g}_{AB} = \eta_{AB} + h_{AB}$, where h is the perturbation whose propagation we want to study. We choose to work in the harmonic gauge in which $\partial^A h_{AB} = \frac{1}{2}\partial_B h_A^A$. Since the extra condition $h_{4\mu} = 0$ is compatible with the equations of motion of (13.68), we are finally left with only $h_{\mu\nu}$ and h_{44} to determine. We then find that

$$\partial_A \partial^A h^\mu_\mu = \partial_A \partial^A h^4_4,$$

as well as

$$\left[\frac{\partial_A \partial^A}{2\kappa_5} + \frac{\delta(y)}{2\kappa_4} \square \right] h_{\mu\nu}(x^\alpha, y) = \left(T_{\mu\nu} - \frac{1}{2} \eta_{\mu\nu} T + \frac{\partial_\mu \partial_\nu h^4_4}{2\kappa_4} \right) \delta(y) + \frac{1}{\kappa_5} \eta_{\mu\nu} \partial_A \partial^A h^\alpha_\alpha,$$

an equation that can be interpreted, on the brane at $y = 0$ as describing the propagation of a massless graviton together with a scalar field. Gravity produced in this way is that of a scalar-tensor theory, including a continuum of massive gravitons.

The graviton propagator on the brane is given by

$$h_{\mu\nu}(k, y = 0) \tilde{T}^{\mu\nu} = \left(\frac{k^2}{2\kappa_4} + \frac{k}{\kappa_5} \right)^{-1} \left(\tilde{T}^{\mu\nu} \tilde{T}_{\mu\nu} - \frac{1}{3} \tilde{T}_\mu^\mu \tilde{T}^\nu_\nu \right),$$

which allows a characteristic length $r_0 = \kappa_5/2\kappa_4$ to be defined, obtained by setting both terms in k^2 and k in the propagator of the same order of magnitude at this scale. For distances $r \ll r_0$, the propagator leads to a gravitational interaction potential scaling as r^{-1} with logarithmic corrections. For $r \gg r_0$, of course, we recover a potential going as r^{-2} , merely pointing out the existence of a fifth dimension. It is interesting to note that here, unlike in Kaluza–Klein models or with warped geometry, there is also an important modification of gravity at large distances. The observational predictions are therefore very different. In particular, such models have been considered to account for an acceleration of the present-time Universe without including any explicit dark-energy contribution (Chapter 12).

13.4 Initial singularity and a bouncing Universe

Even when resorting to an inflationary phase, the standard cosmological model does not solve all of its problems. Phenomenologically, it is a very powerful and efficient paradigm in agreement with observations, but it does not prevent the existence of a singularity in the past. String cosmology pretends to solve this singularity by arguing that corrections to the effective theory become important while the classical description is still valid (note that we still hope that there are no singularities in string theory). This translates into the existence of a maximal curvature, and as a consequence the current phase, including that describing inflation, must have been preceded by another phase. In the string (Jordan) or the Einstein frame, this earlier phase can often be described by a contracting phase. In these scenarios, the Universe bounces as it passes through a minimal scale factor.

13.4.1 Approaching the singularity

The fact that the Friedmann–Lemaître space-time provides a good description of the later phases of the Universe in no way implies that this space-time describes correctly the primordial phases of the Universe. As seen in the example of anisotropic homogeneous spaces (Section 3.6.3, Chapter 3), the shear decreases in time and could hence have dominated in the past, when the Universe was close to the initial singularity.

13.4.1.1 Bianchi I Universe

The Bianchi I solution describes a homogeneous space-time with Euclidean but anisotropic spatial sections. Its metric takes the form (3.121). As shown by the study of the solution of Section 3.6.3 from Chapter 3, the shear dominates the solution when $t \rightarrow 0$ (3.122). It always dominates the matter contribution in the primordial stages and we can hence study the nature of the singularity by neglecting the matter contribution.

The solution (3.122) shows that $S \propto t^{1/3} \rightarrow 0$. The singularity always exists and is a real singularity (the geometrical invariants of the metric always diverge). However, unlike the case of a Friedmann–Lemaître space-time, two scale factors tend to 0 while one dimension keeps expanding.

These conclusions on the nature of the approach to a singularity depend on the matter content of the Universe, on the space-time symmetries and on the theory of gravity at high energy.

13.4.1.2 Bianchi IX universe

We now describe the original study of the approach to the singularity such as initially presented in Ref. [28].

The Bianchi IX solution is analogous to that of Bianchi I except that it describes an anisotropic homogeneous universe with positive curvature spatial sections. In a coordinates system in which the structure constants (defined in Section 3.6.3 of Chapter 3) are $N^{11} = N^{22} = N^{33} = 1$ (see Section 118 of Ref. [29] for its construction), the spatial metric involves three scale factors $X(t)$, $Y(t)$ and $Z(t)$. When $X = Y = Z$, this space reduces to a Friedmann–Lemaître space-time with spherical spatial sections. As in the case of a Bianchi I universe, the shear dominates in the neighbourhood of the singularity so that the matter contribution is negligible. The Einstein equations then reduce to

$$\frac{d(\dot{X}YZ)}{dt} = \frac{1}{2XYZ} \left[(Y^2 - Z^2)^2 - X^4 \right], \quad (13.69)$$

$$\frac{\ddot{X}}{X} + \frac{\ddot{Y}}{Y} + \frac{\ddot{Z}}{Z} = 0, \quad (13.70)$$

corresponding, respectively, to the equations $R_1^1 = 0$ and $R_0^0 = 0$. There are two other similar Einstein equations obtained by circular permutation between X , Y and Z from (13.69). It is useful to introduce a new time variable τ defined by

$$dt = (XYZ)d\tau, \quad (13.71)$$

as well as new functions α , β and γ by

$$X = e^\alpha, \quad Y = e^\beta, \quad Z = e^\gamma, \quad (13.72)$$

in terms of which the Einstein equations now reduce to

$$2\alpha'' = \left[(Y^2 - Z^2)^2 - X^4 \right], \quad (13.73)$$

$$\alpha'' + \beta'' + \gamma'' = 2(\alpha'\beta' + \alpha'\gamma' + \beta'\gamma'), \quad (13.74)$$

with, as above, two equations obtained from (13.73) via circular permutation of α , β and γ .

To study the dynamics of this universe in the neighbourhood of the singularity, we note first that the second term in (13.69) is negligible, which amounts to saying that the curvature term is negligible. This means that the Bianchi IX space is very close to a Bianchi I universe, simply because the curvature is negligible with respect to the shear, so that the scale factors behave in a way identical to that of the Kasner metric, namely

$$X \propto t^{p_x}, \quad Y \propto t^{p_y}, \quad Z \propto t^{p_z}, \quad (13.75)$$

the p_i exponents satisfy the two constraints

$$p_x + p_y + p_z = p_x^2 + p_y^2 + p_z^2 = 1. \quad (13.76)$$

These three exponents are a permutation of (p_1, p_2, p_3) with $p_1 \leq 0 \leq p_2 < p_3$. It can be explicitly expressed as

$$p_1(u) = -\frac{u}{\sigma(u)}, \quad p_2(u) = \frac{u+1}{\sigma(u)}, \quad p_3(u) = \frac{u^2+u}{\sigma(u)}, \quad (13.77)$$

with $\sigma(u) = 1+u+u^2$. Varying u between 1 and $+\infty$ provides a complete representation of all possible triplets (p_1, p_2, p_3) . Moreover, one can check that the transformation $u \rightarrow 1/u$ keeps p_1 unchanged and exchanges p_2 and p_3 . Finally, note that the solution (13.75) implies that

$$XYZ \propto t, \quad \tau = \ln t + C, \quad (13.78)$$

where C is a constant.

Let us now consider an epoch during which the Universe is well described by a Kasner solution with, for instance, $p_x = p_1$. When the system evolves towards the singularity, i.e. when $t \rightarrow 0$, Y and Z tend to 0 while X increases. So it can no longer be neglected in (13.69), which then takes the form

$$\alpha'' = -2e^{4\alpha}, \quad \beta'' = 2e^{4\alpha}, \quad \gamma'' = 2e^{4\alpha}. \quad (13.79)$$

These equations are analogous to that of a massive particle evolving in an exponential potential. It is a good approximation to replace such a steep potential by a wall perpendicular to the axis x . What happens to this particle is then clear: when it reaches the wall with a speed $\alpha' = p_1$, it bounces back with a speed $\alpha' = -p_1$. After the bounce, the space-time is described by a new Kasner solution but with different exponents. Equations (13.79) imply that β' goes from p_2 to $p_2 + 2p_1$ and γ' from p_3 to $p_3 + 2p_1$. The relation between t and τ is thus modified into $t = e^{(1+2p_1)\tau}$. We end up with new exponents given by

$$p_x \rightarrow p_x = -\frac{p_1}{1+2p_1}, \quad p_y \rightarrow p_y = \frac{p_2 + 2p_1}{1+2p_1}, \quad p_z \rightarrow p_z = \frac{p_3 + 2p_1}{1+2p_1}, \quad (13.80)$$

the sum of which remains unity, as expected. So the curvature term can be considered as acting during a very short period of time compared to the Kasner eras and induces the modification discussed above in the exponents (p_x, p_y, p_z) so that we go from one Kasner era to another. The negative exponent and the smallest of the positive

exponents change sign during the bounce. The scale factor associated with the negative exponent before the bounce passes through a minimum, while those for which the exponent becomes negative passes through a maximum. Meanwhile, the third scale factor keeps decreasing.

The singularity is therefore approached via a series of oscillations with the volume decreasing approximatively as t . The Kasner phases are shorter and shorter and the process takes the form of a random process, where the number of oscillations between any time $t > 0$ and the singularity is always infinite. The nature of the singularity approach is thus radically different from the case of a Friedmann–Lemaître space-time but the singularity still exists.

The series of Kasner phases can be obtained from that of the parameter (u_n) coming in the definitions (13.77) of the exponents. This series is defined by the recursive relation

$$u_{n+1} = u_n - 1 \quad (2 \leq u_n), \quad u_{n+1} = \frac{1}{u_n - 1} \quad (1 \leq u_n \leq 2). \quad (13.81)$$

Depending on the value of u_1 , it can be periodic or chaotic.

13.4.1.3 Approaching the singularity

Approaching the singularity can be very different depending on the model for the Universe we assume to begin with. The fact that our Universe seems to be well described by a Friedmann–Lemaître Universe now therefore does not allow us to draw any conclusions about the nature and dynamics of its primordial phase. In the previous example, the cosmological solution close to the singularity is described by a chaotic sequence [30] of Kasner solutions, even though it could lead to a regular Friedmann–Lemaître Universe at large times, i.e. very far from the singularity.

The fact that there should exist a singularity can be proven independently of the space-time symmetries using general theorems on singularities, together with general properties of the energy-momentum tensor [31]. As shown, in general, by the Raychaudhuri equation (1.73), if the vorticity and the acceleration of the geodesics flows vanish, the Universe is expanding and the strong energy condition is satisfied, then there must exist a singularity in the past.

Of course, when the curvature of the Universe increases as the singularity is approached, one expects that modifications, induced because of the supposed quantum nature of gravity at scales close to the Planck energy, can change these conclusions. Approaching the singularity in the context of string theory was studied in detail in Ref. [32]. It was found that the chaotic behaviour always exists and that it reveals the underlying structure of the theory.

In what follows, we will focus on models in which the singularity is avoided.

13.4.2 The pre-Big-Bang scenario

The so-called pre-Big-Bang solution is radically different from inflation. It assumes that the Universe has never gone through a singularity. As a consequence, one can analytically continue the time variable before the beginning of the expansion, now merely referred to as the Big Bang, and ask questions about the phase that precedes

it. The question is then to know if it is possible that this phase leads to a satisfying initial state for the subsequent evolution, without needing inflation after the Big Bang. This scenario can be implemented, for instance, thanks to the kinetic term of the dilaton, that can cause a phase similar to inflation, but before the Big Bang. Note that both ideas do not exclude each other: one can perfectly imagine a model in which the pre-Big-Bang contribution is used to reach a post-Big-Bang phase in an acceptable initial state to induce a subsequent phase of inflation. In what follows, we will assume that there is no post-Big-Bang inflationary phase, so that the effects of the pre-Big-Bang evolution are not erased. This allows us to evaluate whether they are accessible observationally [33].

The pre-Big-Bang model belongs to this category of scenarios that suffers from vacuum instability and is thus plagued with tremendous difficulties. However, besides its historical interest, it also exemplifies a means of implementing string ideas such as the dilaton and dualities, in a cosmological setting.

13.4.2.1 Superinflation and accelerated contraction

Since the primordial singularity is from now on replaced by a large-curvature regime, the pre-Big-Bang phase must therefore evolve towards large curvature, while keeping the property of having $K/a^2 \ll H^2$, i.e. $aH \rightarrow 0$, so as to be able to solve the curvature problem. In an inflationary phase, the power-law or exponential accelerated expansion gives this result. Another way to obtain it is to consider a phase of power-law expansion, $a \sim t^\beta$, ending at $t \rightarrow 0^-$, which requires $\beta < 1$. For $\beta < 0$, we find that \dot{a} , \ddot{a} and \dot{H} are positive, it is a phase of accelerated expansion, called superinflation. For $0 < \beta < 1$, all these quantities are now negative, and it is then an accelerated contraction. In the two later cases, the 4-dimensional curvature increases, which is exactly what we were looking for in a pre-Big-Bang phase.

We now consider the action (13.36) in D dimensions, in which we neglect the Kalb–Ramond term. For a cosmological configuration with no spatial curvature for which $g_{AB} = [N^2(t), a(t)\delta_{ij}]$, after integrating by parts, we obtain the action⁴

$$S_{\text{Jordan}} = \frac{1}{2\kappa_D} \int d^Dx a^{D-1} N^{-1} e^{-\phi} \left[(D-1)(D-2) H^2 + \dot{\phi}^2 - 2(D-1) H \dot{\phi} \right]. \quad (13.82)$$

In the Einstein frame, it takes the form

$$S_{\text{Einstein}} = \frac{1}{2\kappa_D} \int d^Dx \tilde{a}^{D-1} \tilde{N}^{-1} \left[(D-1)(D-2) H^2 - \frac{1}{2} \dot{\phi}^2 \right], \quad (13.83)$$

from which we can quickly identify the change of frame

$$\tilde{a} = a \exp \left(-\frac{\phi}{D-2} \right), \quad \tilde{N} = N \exp \left(-\frac{\phi}{D-2} \right), \quad \text{and} \quad \tilde{\phi} = \phi \sqrt{\frac{2}{D-2}}. \quad (13.84)$$

⁴Note that plugging the ansatz of the solution before varying the action may be dangerous and may lead to spurious solutions.

The functions $a(t) \propto (-t)^{-1/\sqrt{D-1}}$ and $e^\phi = (-t)^{-(1+\sqrt{D-1})}$ are solutions of the equations of motion derived from the action (13.82) for $t \rightarrow 0^-$. They represent the case of superinflation described above.

To know what the superinflation solution looks like in the variables of the other representation, we must pick a gauge for the function N . In both frames, we can assume the synchronous gauge, defined as being the one in which, in the string frame variables, $N = 1$. If we also want to choose this gauge in the Einstein frame, this implies that we take the time coordinate $d\tilde{t} = e^{-\phi/(D-2)}dt$, which, given the solution, can be integrated, into $t \propto \tilde{t}^{(D-2)/(D-1+\sqrt{D-1})}$. The solution, in this new frame and the same gauge is thus $a(t) \propto (-t)^{1/D-1}$ and $e^\phi = (-t)^{-\sqrt{2(D-2)/(D-1)}}$, still in the limit $t \rightarrow 0^-$. This new solution represents a case of accelerated contraction, showing that the two possible pre-Big-Bang cases actually only form one, depending on the representation from which we look at it; we should still decide what is physically observable, and we usually assume that it is in the string frame that we should place ourselves for that (see the discussion on the scalar-tensor theories of Chapter 10 for analogous considerations). The solution in both the string and the Einstein frames is depicted in Fig. 13.9.

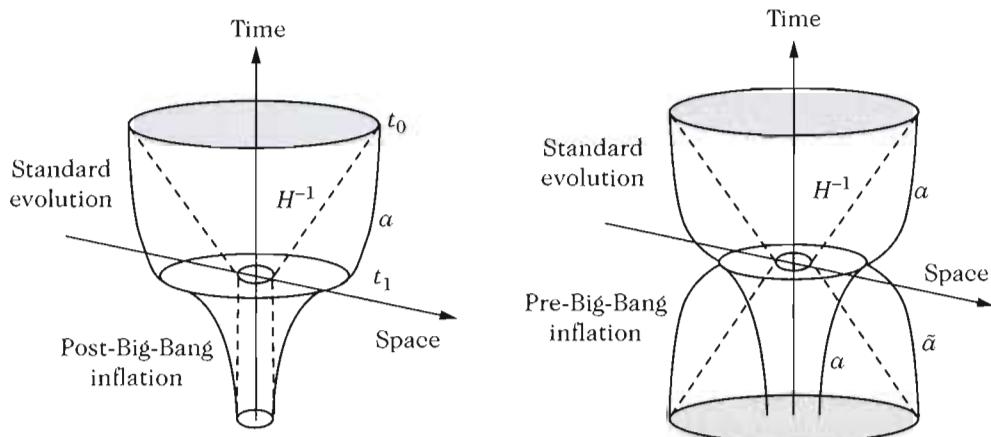


Fig. 13.9 Evolution of the scale factor in the cases of inflation (left) and pre-Big Bang (right). These diagrams show the evolution of the scale factor with time in both situations, as well as the Hubble factor, which is almost constant during the inflationary phase. In the right diagram, we see both the Einstein frame, in which the scale factor goes through a minimum and bounces, and the string frame where there is a superinflationary phase. In the cases where the scale factor is a power law, the horizon is proportional to the Hubble length; the shaded region at the top of each figure representing the actual Hubble radius of the Universe, we see how both models solve the horizon problem differently.

13.4.2.2 Time reversal and duality

Since Einstein equations are invariant under time reversal, we find that if $a(t)$ is a solution, then so is $a(-t)$. This new solution corresponds to a different sign of the expansion rate, so that if one solution describes an expansion, the other describes a contraction.

Considering the action (13.36), still without the Kalb–Ramond term, considering for definiteness a metric of the Bianchi I type [see metric (3.121) of Section 3.6.3 from Chapter 3], with $D - 1$ different scale factors a_i , and setting $\bar{\phi} = \phi - \sum_i \ln a_i$, then the equations of motion are

$$\dot{\bar{\phi}}^2 = \sum_i H_i^2, \quad \dot{H}_i = H_i \dot{\bar{\phi}}, \quad \text{and} \quad 2\ddot{\bar{\phi}} - \dot{\bar{\phi}}^2 = \sum_i H_i^2. \quad (13.85)$$

These equations show two types of symmetries: first the time reversal, now defined more broadly by the transformations $t \rightarrow -t$, $H_i \rightarrow -H_i$, and $\dot{\bar{\phi}} \rightarrow -\dot{\bar{\phi}}$. We also find that if we impose that $\bar{\phi}$ is invariant, then the transformation $a_i \rightarrow a_i^{-1}$ keeps the system invariant. This new symmetry, called *scale factor duality* can be used to obtain other solutions from a single one. Indeed, if $\{a_1, \dots, a_{D-1}, \phi\}$ is a solution, then $\left\{a_1^{-1}, \dots, a_k^{-1}, a_{k+1}, \dots, a_{D-1}, \phi - 2 \sum_{i=1}^k \ln a_i\right\}$ is also a solution. In the isotropic case, $a_i = a$, and we have a symmetry under the transformations $a \rightarrow a^{-1}$ and $\phi \rightarrow \phi - 2(D-1) \ln a$, which here again transforms a contracting phase to an expanding phase since H changes sign during this transformation. As a consequence, for every scale factor (and dilaton) solution, one can actually associate four solutions, $a(t)$, $a(-t)$, $a^{-1}(t)$ and $a^{-1}(-t)$, two of which are expanding and two of which are contracting.

The fact that there are such transformations between the solutions provides a hint that the pre-Big-Bang expanding phase could have been preceded by a dual phase, either contracting or expanding, depending on which frame we work in. Connecting one to the other then gives a complete solution supposed to describe the cosmological evolution for all times.

An interesting solution that illustrates the general properties of these models is given by $a = (t/t_0)^{1/\sqrt{D-1}}$ and $\bar{\phi} = -\ln(t/t_0)$ with its transformations. These four branches, expansion and contraction, both decelerated or accelerated, are separated by a curvature singularity. If, starting from $\bar{\phi}$, we go back to the definition of the dilaton ϕ , then we can easily convince ourselves that if the curvature increases, then the dilaton can only increase for the expanding phases. During this phase, the string coupling constant $g_s = e^\phi$ also increases, so that the cosmological solution derives away from the solution that is valid perturbatively: in the pre-Big-Bang framework, the initial configuration is that of a very weakly coupled string vacuum.

13.4.2.3 Dilaton potential and matter coupling

One often adds a potential term $V(\phi)$ for the dilaton. Although such a term is absent at the perturbative level, there are good reasons to suppose that it can appear as a non-perturbative effect from string theory. Besides, such a term breaks the symmetries we have just described, unless V only depends on $\bar{\phi}$ [see above (13.85)] instead of ϕ .

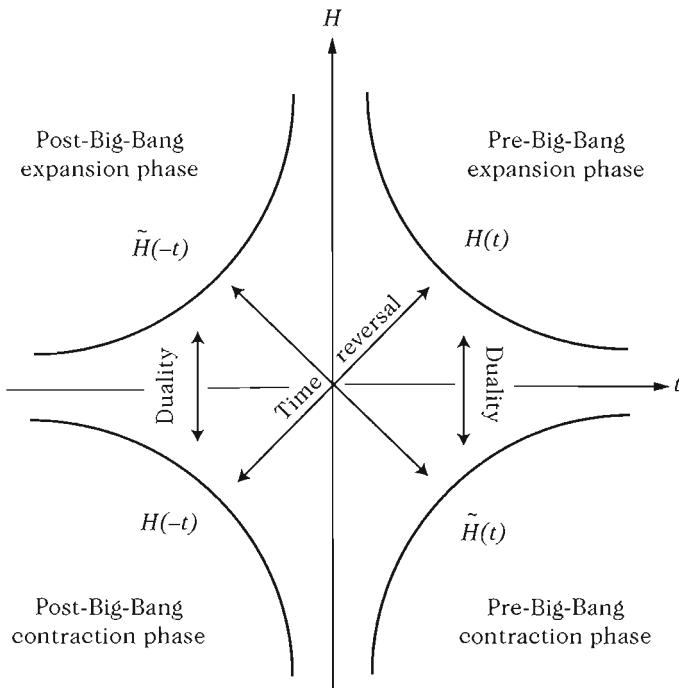


Fig. 13.10 String cosmology gives four solutions for a given scale factor, related to each other by time-reversal $t \rightarrow -t$ or scale-factor duality $a \rightarrow 1/a$. The relations between the different configurations, duality or time-reversal, are given explicitly for each set of two configurations.

For a potential endowed with reasonable properties, one can still find cosmological trajectories akin to those described by Fig. 13.11. It is a generic feature of these kind of scenarios to start with an initial condition in which the dilaton is in a perturbative vacuum state. It is called the asymptotic triviality condition.

The time-reversal and duality symmetries are still present in the case of coupling with matter, as long as this coupling is only realized with gravity and not the dilaton. There are still four branches, related by the exchange symmetries, but these are unfortunately not that simple to realize explicitly. For instance, in the case of a perfect fluid, the equation of state should sometimes be reversed $w \rightarrow -w$, although the resulting model is pathology free provided the initial one was. We note, furthermore, that for a radiation-dominated Universe in $D = 4$ dimensions, the decelerated expansion solution we obtain is indeed that of general relativity $a \propto \sqrt{t}$, for which the dilaton does not depend on time.

Precise models are discussed in detail in Ref. [33]. They share the common property to start from an asymptotically trivial solution. Such a condition is similar to requesting an asymptotically flat space-time in computations of spherically symmetric solutions in general relativity, and then to evolve this configuration until the non-perturbative regime through a gravitational collapse. In general relativity, this eventually produces a singularity, which is avoided here only by the non-perturbative cor-

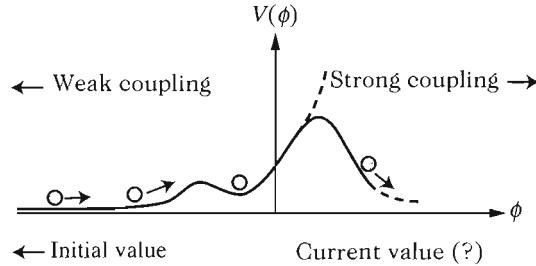


Fig. 13.11 In a model where the dilaton has a potential term induced by non-perturbative corrections of string theory, its trajectory can be the following: if one sets ϕ initially in the perturbative vacuum, namely $\phi \rightarrow -\infty$ where the potential is very weak, then it will move up hill along the potential until it reaches the strong curvature regime. After it has passed through, it then plummets back again to an expanding phase with decreasing curvature. In fact, in this approach, the dilaton can be either stuck at a minimum of its potential or roll down the decreasing part of the potential towards $\phi \rightarrow \infty$.

rections of string theory. By a duality transformation, the Universe then turns into an expanding phase similar to the post-Big-Bang.

13.4.2.4 Perturbations

Pre-Big-Bang cosmology becomes predictive as soon as the gauge invariant perturbations it can produce are taken into account, in order to be able to compare the expected spectra with observations. First, we show that these spectra are identical whether they are computed in the string or in the Einstein frame.

Let us take, for instance, a tensor perturbation on the metric (in synchronous gauge)

$$ds^2 = -dt^2 + a^2(t)\delta_{ij}dx^i dx^j + b^2(t)\delta_{mn}dx^m dx^n, \quad (13.86)$$

where $i, j = 1, \dots, d$, $m, n = d+1, d+n$ and the total space-time dimension is $D = 1 + d + n$. Tensor perturbations $\delta g_{AB} = h_{AB}$ with $\nabla_B h_A^B = 0 = h_A^A$ satisfy, for the dimensions whose dynamics is given by the scale factor a and for each polarization mode h

$$h'' + \left[(d-1) \frac{a'}{a} + n \frac{b'}{b} - \phi' \right] h' - \nabla^2 h = 0, \quad (13.87)$$

in the string frame, as a function of the conformal time $d\eta = dt/a$ and

$$\tilde{h}'' + \left[(d-1) \frac{\tilde{a}'}{\tilde{a}} + n \frac{\tilde{b}'}{\tilde{b}} \right] \tilde{h}' - \nabla^2 \tilde{h} = 0, \quad (13.88)$$

in the Einstein frame. Both these equations seem different at first sight, but one has to bear in mind that they do not refer to the same observable quantities.

Given the conformal transformation $g_{AB} = e^{2\phi/(d-1)} \tilde{g}_{AB}$ and the fact that both conformal times are the same $d\eta = dt/a = d\tilde{t}/\tilde{a} = d\tilde{\eta}$, we have $\tilde{a} = ae^{-\phi/(d+n-1)}$ and $\tilde{b} = be^{-\phi/(d+n-1)}$, so that $(d-1)(\tilde{a}'/\tilde{a}) + n(\tilde{b}'/\tilde{b}) = (d-1)(a'/a) + n(b'/b) - \phi'$

and both perturbation equations are indeed identical. This property is also valid for scalar perturbations, so that one can choose one frame rather than the other for the purposes of computation, with no difference.

The scalar perturbation spectrum one can obtain with the pre-Big-Bang scenario is very different from what is actually observed. One way out of this difficulty is to invoke a curvaton mechanism thanks to which the correct power spectrum can be transferred from another field into the Bardeen potential. Then, neither the amplitude nor the spectral index are constrained, and therefore the model completely lacks any explanatory power. It, however, illustrates the kind of completely different scenarios that can be derived from string theory.

13.4.3 The cyclic scenarios

Another category of models, also incompatible with the data, has been suggested, based on extra dimensions and branes. When discussed side by side with the pre-Big-Bang scenario, it illustrates the amazingly wide variety of possibilities opened up by string theory. Put in a different perspective, this very fact would seem to show that, even in the framework of cosmology, string theory does not yet have any firm and definite prediction. This is not the case though, as taking into account stability requirements [1] strongly limits the possibilities. In fact, apart from the mention of branes, these scenarios hardly have anything in common with string theory itself. They were even argued to require ghosts [35].

The cyclic model of the Universe also postulates the existence of a phase preceding the standard Big Bang. Actually, it consists of a very large number of such phases, our current cosmological era being only one phase amongst many others. Unlike models of inflation, this model postulates that the vacuum energy we currently observe is part of the scenario. But just like inflation, and contrary to what the name ‘cyclic’ might suggest, the Universe it leads to has grown tremendously, although not at once: instead of an inflation phase, our Hubble patch now originates from a very small region of space that expanded after each bounce, the maximum value of the scale factor increasing at each cycle.

We also want to mention the class of so-called *emergent* models whose existence is not directly related to string theory but also has the feature that it avoids the presence of a singularity [36].

13.4.3.1 Ekyrotic part

Let us now concentrate on the so-called ekpyrotic part of the cyclic scenario. One assumes that our Universe is a brane forming a boundary of space, as in the Hořava-Witten model. The other boundary is also a brane of the same dimensionality as ours. Both branes are supposed to be capable of interacting with one another, and in the four-dimensional effective theory, this interaction translates into a scalar field, a function of the distance between the branes (in other words the radion). This field evolves in a potential assumed to arise non-perturbatively from string theory, and is represented in Fig. 13.12. It is not clear whether such a brane interaction potential is actually even possible in string theory, but for the most part, the model is said ‘not to depend crucially on the details’.

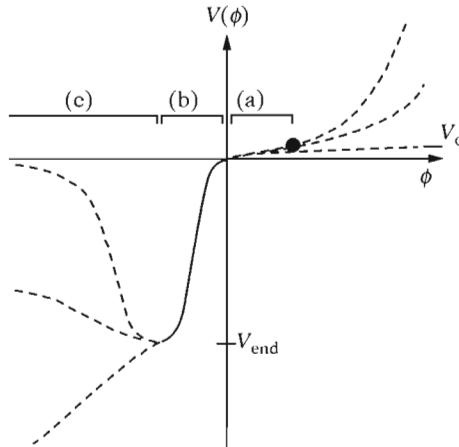


Fig. 13.12 Effective potential of the four-dimensional effective scalar field, thanks to which the ekpyrotic model describes most phases of the Universe. In region (a), the vacuum energy slowly disappears, until the end of the accelerated expansion period we currently know. In region (b), the potential is negative and exponential. Scale invariant perturbations are produced in this phase. Finally, in region (c), the Universe contracts until a singularity, and a new cycle starts. For numerical computations, a form sharing all the required characteristics for such a potential is $V(\phi) = V_0 \left(1 - e^{-\phi/m_1}\right) \exp\left(-e^{-\phi/m_2}\right)$.

The four-dimensional effective theory is defined by the action

$$S_{\text{ekp}} = \int_{\mathcal{M}} d^4x \left[\frac{1}{2\kappa} R - \frac{1}{2} (\partial\phi)^2 - V(\phi) \right], \quad (13.89)$$

where the potential, for the phase of interest, should have the shape shown in (b) of Fig. 13.12, which can be well approximated by a negative exponential, namely

$$V(\phi) = -V_i \exp\left[-\frac{4\sqrt{\pi\gamma}}{M_p} (\phi - \phi_i)\right]. \quad (13.90)$$

This is, up to a sign, merely a power-law inflationary potential. In (13.90), V_i , ϕ_i are constants, just like the function $\gamma = 1 - \mathcal{H}'/\mathcal{H}^2$, when the scale factor is a solution of the Einstein equations.

After the phase dominated by this potential, when both branes are very close to one another and the collision is about to happen, the potential rapidly tends to zero, so that the collision occurs during a phase dominated by the kinetic energy of the scalar field. This is equivalent to a fluid whose equation of state is $w = 1$, and we then observe gravitational collapse followed by a bounce after going through a singularity, and the scale factor in conformal time behaves as $a \propto |\eta|^{1/2}$. Instead of solving the singularity problem of standard cosmology, this kind of model involved a whole series of singularities!

Although the singularity obtained at the beginning of each cycle is not removable, it has been argued that it is sufficiently mild to be incorporated into a consistent

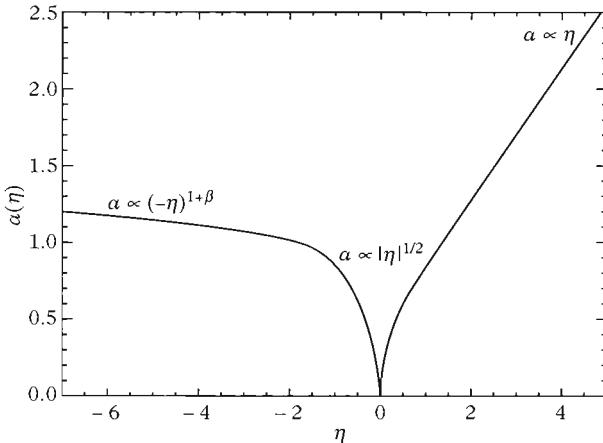


Fig. 13.13 In the ekpyrotic scenario, the scale factor first goes through a slowly contracting phase, $a \propto (-\eta)^p$, $p \ll 1$, and being a scalar field with an exponential potential, it hence produces an almost scale invariant power spectrum. Then, when both branes get closer, the dynamics is dominated by the scalar field kinetic part, until the collision, which in four dimensions is seen as a singularity, and produces entropy. Both branes then start moving away from one another again, producing an expanding phase dominated by the radiation produced during the singularity.

framework. It has also been argued that the perturbation could propagate smoothly across this singularity. In fact, the very existence of the singularity is essential to end up with a scale invariant spectrum for the scalar perturbations. Unfortunately, for tensor perturbations not to diverge, therefore breaking down the perturbation theory, it would be necessary to smooth the singularity itself. As a result, the model ought to be singular and regular at the same time.

13.4.3.2 Power laws and the primordial spectrum

With an exponential potential, the solution for the scale factor is $a = \ell_0(-\eta)^{1+\beta}$ for $\eta < 0$ (we place the origin of time, $\eta = 0$, at the moment of the singularity produced by the collision), with $\gamma = (2 + \beta)/(1 + \beta)$. The solution for the scalar field is given by $\phi = \phi_i + \frac{1}{2}M_p\sqrt{\gamma/2}(1 + \beta)\ln(-\eta)$, and since the Hubble parameter is $aH = (1 + \beta)/\eta$, the expansion follows a slow contraction as long as $0 < \beta + 1 \ll 1$.

Since there is only one scalar field dominating the dynamics, the perturbation equation is (see Chapter 8)

$$v'' + \left[k^2 - \frac{(a\sqrt{\gamma})''}{(a\sqrt{\gamma})} \right] v = 0, \quad (13.91)$$

where the Bardeen potential is expressed in terms of v through the relation

$$\Phi = \frac{\mathcal{H}\gamma}{2k^2} \left(\frac{v}{a\sqrt{\gamma}} \right)'.$$

We then set the initial condition in a Bunch–Davies vacuum (see Chapter 8), which in this case amounts to setting

$$\lim_{k/(aH) \rightarrow +\infty} v = -\frac{4\sqrt{\pi}}{M_p} \frac{e^{-ik(\eta-\eta_i)}}{\sqrt{2k}}, \quad (13.92)$$

and we find the exact solution $v = (k\eta)^{1/2} [A_1(k)J_{\beta+1/2}(k\eta) + A_2(k)J_{-(\beta+1/2)}(k\eta)]$, out of which the Bardeen potential is expressed as

$$\Phi(\eta) = -\frac{\mathcal{H}\sqrt{\gamma}}{2ka} (k\eta)^{1/2} [A_1(k)J_{\beta+3/2}(k\eta) - A_2(k)J_{-(\beta+3/2)}(k\eta)], \quad (13.93)$$

with

$$A_1(k) = \frac{\pi\sqrt{8}}{M_p \cos\beta\pi} \frac{e^{i(k\eta_i - \pi\beta/2)}}{\sqrt{2k}}, \quad \text{and} \quad A_2(k) = -iA_1 e^{i\pi\beta}. \quad (13.94)$$

Taking the limit $k\eta \rightarrow 0$, we find

$$\Phi(\eta) = -\frac{\bar{A}_2(k)}{2k^2} \frac{1+\beta}{\ell_0^2} (-\eta)^{-3-2\beta} - \frac{\bar{A}_1(k)}{2} \frac{2+\beta}{3+2\beta}. \quad (13.95)$$

with

$$\frac{\bar{A}_2(k)}{k^2} \sim k^{-\beta-5/2}, \quad \bar{A}_1(k) \sim k^{\beta+1/2}. \quad (13.96)$$

As a consequence, for $\beta \simeq -1$, the dominant part of the spectrum (the term in \bar{A}_1) is scale invariant, $\Phi \propto k^{-3/2}/\eta$. Note that the situation is very different from power-law inflation; the latter reproduces the same spectrum for $\beta \sim -2$, but with a term of the type \bar{A}_2 and hence that does not depend on time, unlike the previous one.

13.4.3.3 Matching with the expansion

Once the collision happens, we assume that a large quantity of entropy is produced, in the form of radiation and matter, so that the model directly enters in a radiation phase. The perturbation modes are then those of a hydrodynamical phase, these modes being the ones that will mark the fluctuations in the cosmic microwave background. It is therefore important to know how these modes are expressed in terms of the modes in the contracting phase.

An hypothesis is then made according to which the bounce only affects the modes by mixing them. In practice, this means that the dominant D_\pm and subdominant S_\pm modes before (–) and after (+) the bounce are assumed to be related by a transition matrix

$$\begin{pmatrix} D_+ \\ S_+ \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \begin{pmatrix} D_- \\ S_- \end{pmatrix}, \quad (13.97)$$

and we assume the matrix elements to be independent of k . Note that this hypothesis can be justified to some extent for different cases, but, in general, there is no argument that prevents any form of k -mode mixing. As a consequence, with a scale invariant part in the contracting phase, once modes have been mixed at the bounce, the spectrum of the perturbation in the expanding phase should acquire a scale invariant component in the ekpyrotic scenario.

13.4.3.4 End of a cycle

After the collision phase during which the potential vanishes, we assume that the scalar field restarts from a small value of the potential. This phase is first dominated by radiation, then by matter and then finally again by the scalar field potential. At this time, provided one has set the potential to a very small value, one should observe an accelerated expansion with a very small cosmological constant, making the model consistent with current observations. During this period, all the classical perturbations (galaxies and other non-linear structures) are completely washed out so that on a large enough volume, which will serve as a basis for the next cycle, space has become homogeneous and isotropic. This phase lasts until the potential vanishes again, initiating a new contracting phase, producing new scale invariant perturbations. This scenario, as attractive as it may seem, has many unresolved difficulties, such as vacuum instabilities and even ghosts [35]. Among these difficulties, just as for the pre-Big-Bang models, we find the problem of propagating the perturbations through the bouncing phase. Because of such problems however, the actual perturbation spectrum cannot be made scale invariant, and a curvaton mechanism must be invoked, again at the expense of predictive power.

13.4.4 Regular bounce and power spectrum

In order to understand how perturbations can propagate through a bouncing phase, let us assume that it can be regularized, either by string effects, or by other quantum gravity effects that are supposed to describe this period.

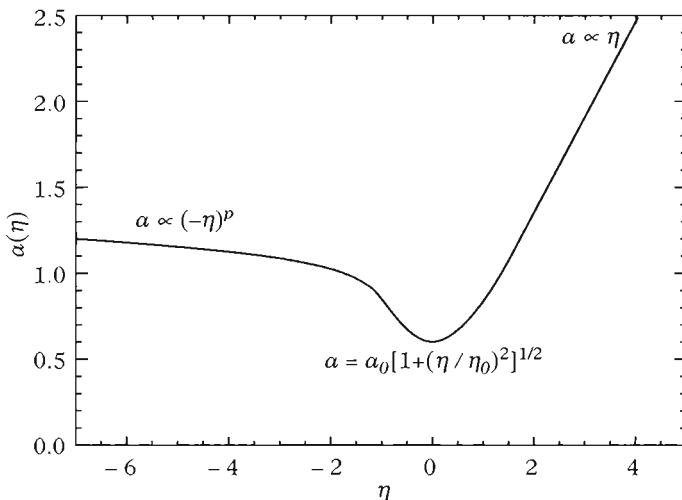


Fig. 13.14 Characteristic form of a regular bounce allowing for the transition between a slow contracting phase, producing a scale invariant component of the spectrum, and a radiation-dominated phase.

13.4.4.1 A class of models

To describe generically the different theories capable of giving rise to a regular bounce such as presented in Fig. 13.14, we split the total action into two components, one geometrical and the other describing the matter content,

$$S_{\text{bounce}} = \int d^4x \sqrt{-g} [\mathcal{L}_{\text{grav}}(R, R_{\mu\nu}, R^\alpha_{\beta\mu\nu}, \dots) + \mathcal{L}_{\text{mat}}(\varphi_i, \psi_a, A_\mu^n, \dots; g_{\mu\nu})], \quad (13.98)$$

where the first term only depends on the metric $g_{\mu\nu}$ and on the tensors we can construct with the Riemann and Ricci tensors (assuming the theory to be free of the pathologies such terms usually produce), and all the possible powers of these objects contracted into scalars, while the second term describes the dynamics of scalar, fermionic vector and other fields, minimally coupled to $g_{\mu\nu}$.

The equations of motion of this action can always be put in the form of the Friedmann equations in flat space-time, namely

$$3\mathcal{H}^2 = \kappa a^2 (\rho_G + \rho_M) \quad \text{and} \quad (2\mathcal{H}' + \mathcal{H}^2) = -\kappa a^2 (p_G + p_M), \quad (13.99)$$

where ρ_M and p_M are the eigenvalues of the matter energy-momentum tensor. We have gathered the other terms, due to the geometrical corrections to Einstein equations under the generic forms ρ_G and p_G . The matter part is then assumed to respect the usual energy conditions, and in particular the weak energy condition, according to which

$$\rho_M + p_M > 0, \quad \text{and} \quad \rho_M > 0,$$

a condition that is, in general, respected at the classical level by all the fields used in cosmology.⁵

The first condition, $\rho + p > 0$, which is the null energy condition (NEC), must be violated at some stage of the evolution of a bounce: we have $\kappa(\rho + p) = 2(\mathcal{H}^2 - \mathcal{H}')/a^2$, where ρ and p represent the sum of each contribution. At the moment of the bounce itself, the scale factor goes through a minimum, i.e. we have $\dot{a} = 0 = a'$, and hence also $\mathcal{H} = 0$. From what precedes, this implies that at the moment of the bounce, $\rho + p < 0$ since, for it to be a minimum, we should necessarily have $\mathcal{H}' > 0$.

We then take the geometric contributions allowing us to realize a bounce. This can be done in several ways, via higher-order terms in the curvature or by non-local terms, all of which arise from string theory. A very simple way to schematize these kinds of terms is to place ourselves in the case where the spatial sections are positively curved. We then have

$$\rho_G = -\frac{3K}{\kappa a^2} \quad \text{and} \quad p_G = \frac{2K}{\kappa a^2},$$

with $K > 0$. This indeed violates the weak-energy condition.

⁵This does not include the category of models with ‘phantom’ matter fields for which the energy is not positive-definite. An example of such a model is a simple scalar field with only a kinetic energy, but one that is negative. This leads to an equation of state $w = 1$ with $\rho < 0$. In this case, we can perfectly realize a completely regular bounce with simply this matter and for instance, radiation [34]. Note that such models are very unstable and must be understood as mere low-energy classical approximations of some instability-free theory.

13.4.4.2 A simple model

Equipped with the positive curvature spatial section term written above, the simplest matter content choice one can make to study a bounce is to take a scalar field whose dynamics stems from a potential that is yet to be determined.

A Universe with closed spatial sections is characterized by two fundamental lengths. The first one is the Hubble length defined by $\ell_H \equiv a^2/a' = a/\dot{a} \equiv H^{-1}$, and the second is the curvature radius of the spatial sections, namely $\ell_C \equiv a/\sqrt{|K|}$. We recover the flat limit as soon as $\ell_C \gg \ell_H$, which can be seen in the equation $|1 - \Omega| = \ell_H^2/\ell_C^2$. For the numerical values, as shown in Chapter 4, one can take, $h = 0.72 \pm 0.05$, giving now a Hubble distance of $\sim 3000h^{-1}$ Mpc $\sim 4.2 \pm 0.2$ Gpc. Moreover, with $\Omega_0 = 1.02 \pm 0.02$, we find a curvature radius that is the scale factor currently measured (Chapter 4), of the order of $a_0 \gtrsim 15h^{-1}$ Gpc, this limit being fixed by the maximal value compatible with the observations of Ω_0 at one standard deviation.

The eigenfunctions $f_n(x^i)$ of the Laplace–Beltrami operator on spatial sections satisfy the relation (see Chapter 12)

$$\Delta f_n = -n(n+2)f_n , \quad (13.100)$$

where $n \in \mathbb{N}$. Note here that the normalization with a dimensioned scale factor $a(\eta)$, and hence dimensionless coordinates (η, x^i) , leads to an operator Δ itself dimensionless, so that its eigenvalues are pure (actually integer) number; with a different convention, i.e. with a dimensionless, we would have had $[\Delta] = L^{-2}$ and an additional factor ℓ_C^{-2} would have appeared on the right-hand side of (13.100).

The mode $n = 0$ corresponds to a homogeneous perturbation, while $n = 1$ is simply a global motion of the 3-sphere centre; both these are therefore gauge modes (see Refs. [4, 5] of Appendix B). In what follows, we will hence only consider modes with $n > 1$. Actually for values of the cosmological parameters in agreement with observational data, we find that for distances of current cosmological interest, namely those ranging between $10^{-2}h^{-1}$ Mpc and 10^3h^{-1} Mpc say, the values of n must be within the limits $30 \lesssim n \lesssim 3 \times 10^6$ for the maximal value of the energy density allowed by the data. For the reasonable value $\Omega_0 \sim 1.01$, we find that n is between 60 and 6×10^6 .

13.4.4.3 Perturbations

When shifting to gauge invariant variables for both the geometry and the matter, the evolution of the Bardeen potential is dictated by (8.140). Defining u by (8.141) and θ by (8.142), it takes the form (8.143). Using the eigenmodes of the Laplacian, u evolves as

$$u'' + \left[n(n+2) - \frac{\theta''}{\theta} - 3K(1 - c_s^2) \right] u = 0 . \quad (13.101)$$

Just as for inflation, to quantize one should define the canonical variable v . For a Universe with non-vanishing curvature, one can extend the definition (8.172) to

$$v = \frac{-a}{\sqrt{1 - 3K \frac{1 - c_s^2}{n(n+2)}}} \left[\chi + \frac{\varphi'}{\mathcal{H}} \Phi - \frac{K\varphi'}{\mathcal{H}^3 \Gamma} \bar{\Phi} \right] , \quad (13.102)$$

with $\Gamma = 1 - \mathcal{H}'/\mathcal{H}^2 + K/\mathcal{H}^2$. This variable then evolves according to the equation

$$v'' + \left[n(n+2) - \frac{z''}{z} - 3K(1 - c_s^2) \right] v = 0, \quad (13.103)$$

which generalizes (8.179), with z now defined by

$$z \equiv \frac{a\varphi'}{\mathcal{H}\sqrt{1 - 3K\frac{1 - c_s^2}{n(n+2)}}}. \quad (13.104)$$

For flat spatial sections, both the variable v and its equation of motion (13.103) reduce to the canonical ones. Figure 13.15 shows the different terms in these equations for a regular bounce whose scale factor is approximated by

$$a(\eta) = a_0 \left[1 + \frac{1}{2} \left(\frac{\eta}{\eta_0} \right)^2 + \delta \left(\frac{\eta}{\eta_0} \right)^3 + \frac{5}{24}(1+\xi) \left(\frac{\eta}{\eta_0} \right)^4 \right], \quad (13.105)$$

where the case presented here is symmetric, i.e. with $\delta = 0$. The form (13.105) represents an expansion around the de Sitter solution, the latter, in the positively curved case, being given by the scale factor $a(t) = a_0 \cosh(t/a_0)$ in terms of the cosmic time, and $a(\eta) = a_0 \sqrt{1 + \tan^2 \eta}$ with the conformal time. Expanding this latter form around $\eta = 0$ to fourth order leads precisely to (13.105) with $\eta_0 \rightarrow 1$ and $\delta, \xi \rightarrow 0$.

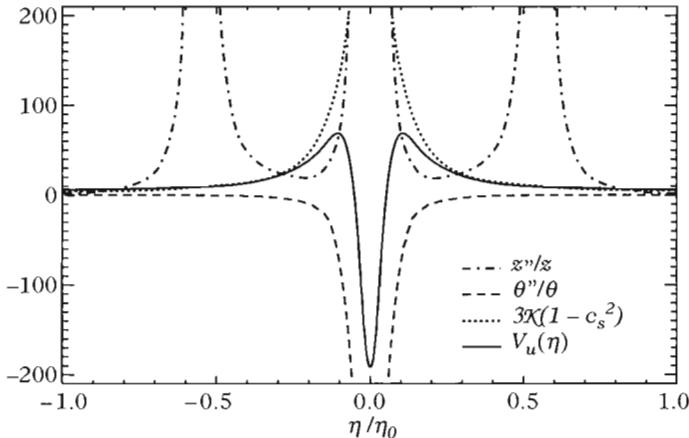


Fig. 13.15 In the case of a regular symmetric bounce, the functions that appear in the equations of motion of the intermediate variables u and v all have divergences, and only the effective potential for u sees these divergences compensating exactly.

The functions z''/z , θ''/θ , $3K(1 - c_s^2)$ and V_u are depicted in Fig. 13.15 as functions of time around the time of the bounce for the parameterization of (13.105). Although most of these functions exhibit divergences either at the NEC violating point

or at the bounce itself, one finds that the potential $V_u(\eta)$ is bounded at all times, the divergences cancelling each other.

13.4.4.4 Propagation

It is important to note that in order for such a bounce to be possible, we should have $\eta_0 \gtrsim 1$, since $\eta_0 < 1$ implies $\phi'^2 < 0$, which is impossible. Without going into too many details, we then find that the bounce will be described in a satisfying way by defining

$$\Upsilon \equiv 1 - \frac{1}{\eta_0^2}, \quad (13.106)$$

and the perturbations will be affected by the bounce in the limit $\Upsilon \rightarrow 0$. In this limit, we then find that there are regimes in the parameter space in which the potential for u can be approximated by

$$V_u(\eta) = -C_\Upsilon \Delta_\Upsilon(\eta), \quad (13.107)$$

where the constant C_Υ is given by $C_\Upsilon \equiv [-5\pi^2\xi/(8\Upsilon)]^{1/2}$ and the function $\Delta_\Upsilon(\eta)$ is a representation of the Dirac distribution, i.e.

$$\lim_{\Upsilon \rightarrow 0} \Delta_\Upsilon(\eta) = \delta(\eta). \quad (13.108)$$

The perturbation equation

$$u'' + [n(n+2) + C_\Upsilon \delta(\eta)]u = 0, \quad (13.109)$$

is then easily solved by taking the matching conditions $[u] = 0$ and $[u'] = -C_\Upsilon u(0)$, which are obtained by integrating (13.109) across the bounce.

In the regions where the potential can be neglected, i.e. before and after the bounce, the solution for u is

$$u_i(n, \eta) = A_i(n)f_i(n, \eta) + B_i(n)g_i(n, \eta), \quad i = \text{I, II}, \quad (13.110)$$

with

$$f_{\text{I,II}}(\eta) = \frac{1}{\sqrt{2\sqrt{n(n+2)}}} e^{-i\sqrt{n(n+2)}\eta}, \quad g_{\text{I,II}}(\eta) = \frac{1}{\sqrt{2\sqrt{n(n+2)}}} e^{i\sqrt{n(n+2)}\eta}, \quad (13.111)$$

the oscillation modes. The relation between the modes before and after is then easily obtained, and we find

$$\begin{pmatrix} A_{\text{II}} \\ B_{\text{II}} \end{pmatrix} = -i\sqrt{\frac{-5\pi^2\xi}{32n(n+2)}} \frac{1}{\Upsilon^{1/2}} \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} A_{\text{I}} \\ B_{\text{I}} \end{pmatrix}, \quad (13.112)$$

which shows that, in this extremely simplified case, the first one we can think of, the modes get mixed, but in addition the spectrum is modified.

Yet another model, based on quantum cosmology and a simple perfect fluid, in this case almost dust-like (i.e. with a very small but non vanishing equation of state,

found of the order $w \leq 8 \times 10^{-4}$ to ensure a possible fit with the amplitude of the power spectrum), has obtained, starting with pure vacuum, a scale invariant spectrum of perturbations. Besides, it is also found that the consistency relation between scalar and tensor modes is modified to $T/S \propto \sqrt{n_s - 1}$ [37]. This kind of model, however, suffers from the requirement that the Universe, at the bounce, should be some 20 orders of magnitude larger than the Planck length.

13.4.4.5 Oscillations in the angular power spectrum

Within the category of bouncing models discussed above, a major prediction is that the oscillation modes $\exp(+ik\eta)$ and $\exp(-ik\eta)$ are excited right after the bounce. As a consequence, we expect oscillations in the primordial spectrum, due to the compression and dilatation modes. The bounce, however, needs a positive curvature to actually occur, and therefore a subsequent phase of inflation. This means that the primordial spectrum is necessarily of the form

$$k^3 |\zeta_{\text{BSR}}|^2 = A(k) \left\{ \alpha + \beta \cos \left[f \left(\frac{k}{k_*} \right) \right] \right\}, \quad (13.113)$$

where the amplitude is a power law $A(k) \sim Ak^n$, getting an extra contribution from the inflation phase, and the function f depends on each model and on the way the bounce is connected to the later phase. The parameters α and β are constants depending on the matter content, the microscopic parameters, and the bounce parameters. For a given model, a characteristic length appears, expressed by k_* , which reflects the scale at which oscillations are produced in the primordial spectrum. It is interesting that a wildly oscillating primordial spectrum having $\alpha = 0$, as illustrated in Fig. 13.16 can fit the otherwise smooth data, as also shown in the figure. This, in fact, comes from the fact that in order to get the observable temperature fluctuation spectrum, one needs to integrate from the surface of last scattering to now, and then to smooth over various directions to obtain the statistical variables we observe.

Once evolved to determine the effect on the observations of the cosmic microwave background fluctuations, such a spectrum induces oscillations superimposed to the standard one. Typical predictions of this category of models are illustrated in Fig. 13.17. Again, and quite surprisingly, these predictions can be made to agree with the observations [38]. And a bounce could also provide new ways [39] of solving the usual cosmological puzzles!

Superstring cosmology [2–4] is, for many reasons, currently an area undergoing rapid expansion. First, there are a multitude of theoretical facts indicating that the string hypothesis could finally allow us to understand gravity at the quantum level. Interestingly, among the recent developments of string theory comes the so-called idea of ‘landscape’ [40], by which one means the representation of all possible vacuum states; the structure of this landscape might provide an anthropic-like solution to the cosmological constant problem [41]: our Universe would then be part of a larger ‘multiverse’,

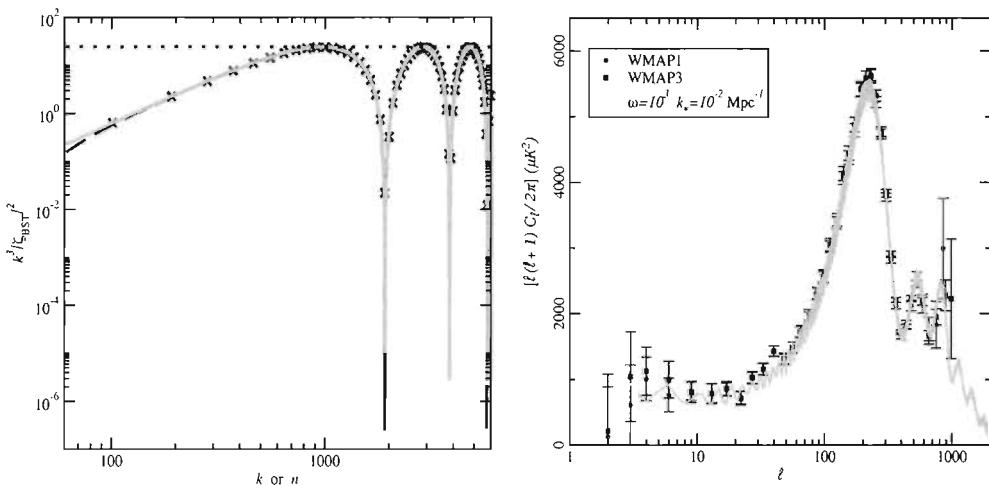


Fig. 13.16 Primordial power spectrum of the curvature perturbation (left) and the corresponding best fit of the angular power spectrum of the cosmic microwave background anisotropies (right) of the WMAP data in a bouncing scenario. Figure from Ref. [38], courtesy C. Ringeval.

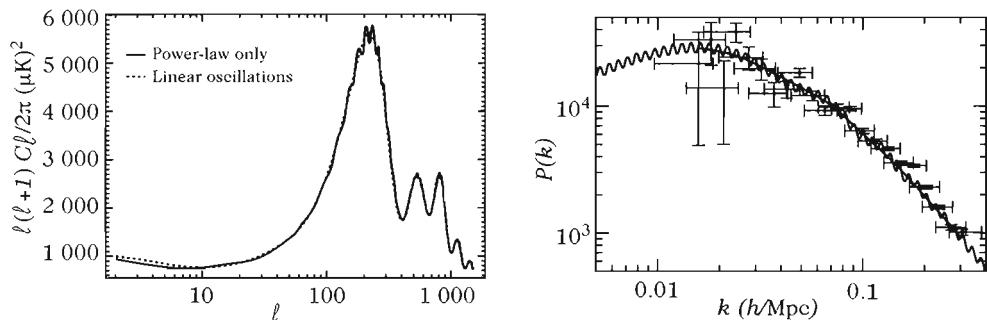


Fig. 13.17 Angular power spectrum of the cosmic microwave background anisotropies (left) and matter power spectrum (right) in a bouncing scenario. The frequency function of the oscillations f in (13.113) is linear with a period k_* much lower than that of Fig. 13.16.

or, in other words, one among many. Interestingly, the fact that this difficulty of string theory could be turned into a virtue had been realized very early in the history of inflation [42].

Apart from this low-energy possibility, the effects of such a theory can unfortunately only occur at considerable energy scales, well beyond anything imaginable at the current state of accelerator technology. It therefore seems natural to consider the only area of physics where such energies can ever have existed. Moreover, since it seems possible that the effects of what happened at these times have left signatures, primor-

dial cosmology can finally be seen as providing experimental input on string theory, a reason why it is important to study all the possible consequences at the cosmological level. While the theoretical developments are quickly providing new ideas and open up further intriguing and fascinating new possibilities, future observational projects may reveal some hints on the ultimate behaviour of Nature!

References

- [1] S. KACHRU, R. KALLOSH, A. LINDE and S. P. TRIVEDI, ‘De Sitter vacua in string theory’, *Phys. Rev. D* **68**, 046005, 2003.
- [2] L. MCALLISTER and E. SILVERSTEIN, ‘String Cosmology: A Review’, *Gen. Rel. Grav.* **40**, 565, 2008.
- [3] D. BAUMANN and L. MCALLISTER, ‘Advances in Inflation in String Theory’, arXiv:0901.0265 [hep-th], 2009.
- [4] R. KALLOSH, ‘Towards string cosmology’, *Prog. Theor. Phys. Suppl.* **163**, 323, 2006.
- [5] T. DAMOUR and A. POLYAKOV, ‘The string dilaton and a least coupling principle’, *Nucl. Phys. B* **423**, 532, 1994.
- [6] G. ESPOSITO-FARÈSE, *Théorie de Kaluza–Klein et gravitation quantique*, PhD thesis from Aix-Marseilles II university, 1989.
- [7] J. M. OVERDUIN and P. S. WESSON, ‘Kaluza–Klein Gravity’, *Phys. Rep.* **283**, 303, 1997.
- [8] B. ZWIEBACH, *A first course in string theory*, Cambridge University Press, 2004; J. POLCHINSKI, *String theory*, Cambridge University Press, 1998; M. B. GREEN, J. H. SCHWARZ and E. WITTEN, *Superstring theory*, Cambridge University Press, 1987.
- [9] Refer to the website of Sunil Mukhi,
<http://theory.tifr.res.in/~mukhi/Physics/string2.html>.
- [10] R. BRANDENBERGER and C. VAFA, ‘Superstrings in the early Universe’, *Nucl. Phys. B* **316**, 391, 1989.
- [11] S. ALEXANDER, R. BRANDENBERGER and D. EASSON, ‘Brane gases in the early Universe’, *Phys. Rev. D* **62**, 103509, 2000.
- [12] R. MAARTENS, ‘Brane-world gravity’, *Living Rev. Relativity* **7**, 7, 2004.
- [13] D. LANGLOIS, ‘Brane cosmology: an introduction’, *Prog. Theor. Phys. Suppl.* **148**, 181, 2003.
- [14] P. BRAX and C. VAN DE BRUCK, ‘Cosmology and brane worlds: a review’, *Class. Quant. Grav.* **20**, R201, 2003.
- [15] P. BRAX, C. VAN DE BRUCK, and A.-C. DAVIS, ‘Brane world cosmology’, *Rept. Prog. Phys.* **67**, 2183, 2004.
- [16] R. DICK, ‘Brane worlds’, *Class. Quant. Grav.* **18**, R1, 2001.
- [17] P. HOŘAVA and E. WITTEN, ‘Heterotic and type I string dynamics from eleven dimensions’, *Nucl. Phys. B* **460**, 506, 1996.
- [18] N. ARKANI-HAMED, S. DIMOPOULOS and G. DVALI, ‘The hierarchy problem and new dimensions at a millimeter’, *Phys. Lett. B* **429**, 263, 1998.
- [19] R. M. WALD, *General relativity*, University of Chicago Press, 1984.

- [20] P. BINÉTRUY, C. DEFFAYET and D. LANGLOIS, ‘Non-conventional cosmology from a brane Universe’, *Nucl. Phys. B* **565**, 269, 2000.
- [21] J. E. LIDSEY, ‘Inflation and braneworlds’, *Lect. Notes Phys.* **646**, 357, 2004; J. E. LIDSEY, D. WANDS and E. J. COPELAND, ‘Superstring cosmology’, *Phys. Rep.* **337**, 343, 2000.
- [22] R. A. BATTYE, B. CARTER, A. MENNIM and J.-P. UZAN, ‘Einstein equations for an asymmetric brane-world’, *Phys. Rev. D* **64**, 124007, 2001.
- [23] N. DERUELLE and T. DOLEZEL, ‘Brane versus shell cosmologies in Einstein and Einstein–Gauss–Bonnet theories’, *Phys. Rev. D* **62**, 103502, 2000.
- [24] C. CHARMOUSIS and R. ZEGERS, ‘Matching conditions for a brane of arbitrary codimension’, *JHEP* **0508**, 075, 2005.
- [25] A. MENNIM and R. A. BATTYE, ‘Cosmological expansion on a dilatonic brane-world’, *Class. Quant. Grav.* **18**, 2171, 2001.
- [26] C. RINGEVÄL, P. PETER and J.-P. UZAN, ‘Localization of massive fermions on the brane’, *Phys. Rev. D* **65**, 044016, 2002.
- [27] G. DVALI, G. GABADADZE and M. PORRATI, ‘4D gravity on a brane in 5D Minkowski space’, *Phys. Lett. B* **485**, 208, 2000.
- [28] V. BELINSKY, E. LIFCHITZ and I. KHALATNIKOV, ‘Oscillatory approach to a singular point in the relativistic cosmology’, *Adv. Phys.* **19**, 525, 1970; *Sov. Phys. JETP* **35**, 383, 1972; *Adv. Phys.* **31**, 639, 1982.
- [29] L. LANDAU and E. LIFSCHITZ, *Course of theoretical physics*, Volume 2, Pergamon Press, 1976.
- [30] E. LIFCHITZ, I. LIFCHITZ and I. KHALATNIKOV, ‘Asymptotic analysis of oscillatory mode of approach to a singularity in homogeneous cosmological models’, *Sov. Phys. JETP* **32**, 173, 1971.
- [31] S. W. HAWKING and G. F. R. ELLIS, *The large scale structure of space-time*, Cambridge University Press, 1973.
- [32] T. DAMOUR, M. HENNEAUX and H. NICOLAI, ‘Cosmological billiards’, *Class. Quantum Grav.* **20**, R145, 2003.
- [33] M. GASPERINI and G. VENEZIANO, ‘The pre-Big-Bang scenario in string cosmology’, *Phys. Rep.* **373**, 1, 2003.
- [34] P. PETER and N. PINTO-NETO, ‘Primordial perturbations in a non-singular bouncing Universe model’, *Phys. Rev. D* **66**, 063509, 2002.
- [35] R. KALLOSH, J. U. KANG, A. LINDE and V. MUKHANOV, ‘The new ekpyrotic ghost’, *JCAP* **0804**, 018, 2008.
- [36] G.F.R. ELLIS and R. MAARTENS, ‘The Emergent Universe: inflationary cosmology with no singularity’, *Class. Quant. Grav.* **21**, 223, 2004.
- [37] P. PETER, E. PINHO and N. PINTO-NETO, ‘Tensor perturbations in quantum cosmological background’, *JCAP* **07**, 014, 2005; P. PETER, E. PINHO and N. PINTO-NETO, ‘Non-inflationary model with scale invariant cosmological perturbations’, *Phys. Rev. D* **75**, 023516, 2007.
- [38] F. T. FALCIANO, M. LILLEY and P. PETER, ‘A classical bounce: constraints and consequences’, *Phys. Rev. D* **77**, 083513, 2008.
- [39] P. PETER and N. PINTO-NETO, ‘Cosmology without inflation’, *Phys. Rev. D* **78**, 063506, 2008.

- [40] R. BOUSSO and J. POLCHINSKI, ‘Quantization of four-form fluxes and dynamical neutralization of the cosmological constant,’ *JHEP* **06**, 006, 2000; S. B. GIDDINGS, S. KACHRU and J. POLCHINSKI, ‘Hierarchies from fluxes in string compactifications,’ *Phys. Rev. D* **66**, 106006, 2002; L. SUSSKIND, ‘The anthropic landscape of string theory’ arXiv:hep-th/0302219, 2003; M. R. DOUGLAS, ‘The statistics of string / M -theory vacua,’ *JHEP* **05**, 046, 2003.
- [41] Q.-G. HUANG and S.-H. H. TYE, ‘The cosmological constant problem and inflation in the string landscape’, arXiv:0803.0663, 2008 and references therein.
- [42] A. D. LINDE, ‘Eternally existing self-reproducing chaotic inflationary Universe’, *Phys. Lett. B* **175**, 395, 1986.

Appendix A

Numerical values

A.1 Physical constants

| | | |
|----------------------------------|---|---|
| Speed of light | c | $2.997\ 9245\ 8 \times 10^8\ \text{m} \cdot \text{s}^{-1}$ (exact) |
| Newton constant | G_N | $6.674\ 28(67) \times 10^{-11}\ \text{m}^3 \cdot \text{kg}^{-1} \cdot \text{s}^{-2}$ |
| Planck constant | \hbar | $6.626\ 068\ 96(33) \times 10^{-34}\ \text{J} \cdot \text{s}$ |
| Reduced Planck constant | $\hbar = \frac{\hbar}{2\pi}$ | $1.054\ 571\ 628\ (53) \times 10^{-34}\ \text{J} \cdot \text{s}$ $= 6.582\ 118\ 99(16) \times 10^{-22}\ \text{MeV} \cdot \text{s}$ |
| Boltzmann constant | k_B | $1.380\ 650\ 4(24) \times 10^{-23}\ \text{J} \cdot \text{K}^{-1}$ |
| Fermi constant | $\frac{G_F}{(\hbar c)^3}$ | $1.166\ 37(1) \times 10^{-5}\ \text{GeV}^{-2}$ |
| Permeability of the vacuum | μ_0 | $4\pi \times 10^{-7}\ \text{N} \cdot \text{A}^{-2}$ |
| Permittivity of the vacuum | $\epsilon_0 = 1/\mu_0 c^2$ | $8.854\ 187\ 817 \times 10^{-12}\ \text{F} \cdot \text{m}^{-1}$ |
| Fine structure constant | $\alpha = \frac{e^2}{4\pi\epsilon_0\hbar c}$ | $1/137.035999679(94)$ |
| Rydberg constant | $R_\infty = m_e c^2 \alpha^2 / 2$ | $13.605\ 691\ 93(34)\ \text{eV}$ |
| Bohr radius | $a_0 = \hbar/m_e c \alpha$ | $5.2918 \times 10^{-11}\ \text{m}$ |
| Thomson scattering cross-section | $\sigma_T = \frac{e^4}{6\pi\epsilon_0^2 m_e^2 c^4}$ | $6.652\ 46 \times 10^{-29}\ \text{m}^2$ |

A.2 Astrophysical quantities

| | | |
|----------------------------|-------------------------------------|--|
| Solar mass | M_\odot | $1.988\ 44(30) \times 10^{30}\ \text{kg}$ $1.189 \times 10^{57}\ M_P$ |
| Solar (equatorial) radius | R_\odot | $6.961 \times 10^8\ \text{m}$ |
| Solar Schwarzschild radius | $R_{s(\odot)} = 2G_N M_\odot / c^2$ | $2.953\ 250\ 08\ \text{km}$ |
| Solar luminosity | L_\odot | $3.846(8) \times 10^{26}\ \text{W}$ |

A.3 Units

A.3.1 Natural units

When using the international system of units (SI), the laws of physics can become quite intricate, involving so many constants, such as the Planck constant \hbar , the speed of light in vacuum c or the gravitational Newton constant G_N [1, 2].

It turns out that one can choose a more adequate system of units that greatly simplifies calculations. This is achieved by imposing that the numerical value of some constants is unity. In general, and this is what was assumed in this book for the most part (unless specified otherwise), cosmologists use the system of units in which ' $\hbar = c = 1$ ' (meaning that this defines a system of units in which the numerical value of the Planck constant and of the speed of light are equal to 1, hence defining this system of units with respect to the SI), the latter equality implying, for instance, that time and distances are expressed in the same units. In this system, masses and energies are also expressed in the same units, and both are inverse to space and time units. This is summarized below.

$$\boxed{\hbar = c = 1 \iff [L] = [T] = [M^{-1}] = [E^{-1}].} \quad (\text{A.1})$$

Choosing the MeV (10^6 eV) or the GeV (10^9 eV) as the basic unit, we have the conversion factor $1 \text{ GeV} \simeq 1.6 \times 10^{-10} \text{ J}$, and it is usually easy to retrieve the numerical factors in the usual system by using $\hbar c \simeq 200 \text{ MeV} \cdot \text{Fm}$.

In order to completely fix the system of units, one must use a third constant. For instance, one can impose that the numerical value of the Newton constant G_N , equal to the inverse of the Planck mass squared in the above system, be set to unity. In this case, one deals with the so-called Planck units. In this system, time, mass, energy and distance are all dimensionless, meaning that the numbers found are in units of the Planck length, time and mass, namely

$$\ell_P = \sqrt{\frac{\hbar G_N}{c^3}}, \quad t_P = \sqrt{\frac{\hbar G_N}{c^5}}, \quad M_P = \sqrt{\frac{\hbar c}{G_N}}.$$

This allows us to define the Planck temperature as $T_P = \sqrt{\hbar c^5 / (G_N k_B^2)}$. In fact, the Boltzmann constant, which merely relates temperature to energy can be understood as providing a definition for the units of temperature. Therefore, one can also set $k_B = 1$ with no physical consequence. With such units, all the dimensionfull physical constants effectively disappear from the equations.

A.3.2 Conversion factors

| | | |
|-----------------------------|--|---|
| Planck mass | $M_p = \sqrt{\frac{\hbar c}{G_N}}$ | $1.220\ 892(61) \times 10^{19} \text{ GeV}/c^2$ $= 2.176\ 44(11) \times 10^{-8} \text{ kg}$ |
| Planck time | $t_p = \sqrt{\frac{\hbar G_N}{c^5}}$ | $5.391\ 24(27) \times 10^{-44} \text{ s}$ |
| Planck length | $\ell_p = \sqrt{\frac{\hbar G_N}{c^3}}$ | $1.616\ 252(81) \times 10^{-35} \text{ m}$ |
| Planck temperature | $T_p = \sqrt{\frac{\hbar c^5}{G_N k_B}}$ | $1.416\ 785(71) \times 10^{32} \text{ K}$ |
| Planck \leftrightarrow SI | 1 GeV | $1.602\ 2 \times 10^{-10} \text{ J}$ $1.160\ 5 \times 10^{13} \text{ K}$ $1.782\ 7 \times 10^{-27} \text{ kg}$ 1 GeV $^{-1}$ $1.973\ 3 \times 10^{-16} \text{ m}$ $6.652\ 2 \times 10^{-25} \text{ s}$ |
| 1 cm | = | $5.068 \times 10^{13} \text{ GeV}^{-1} \hbar$ |
| 1 s | = | $1.519 \times 10^{24} \text{ GeV}^{-1} \hbar/c$ |
| 1 g | = | $5.608 \times 10^{23} \text{ GeV} \hbar/c^2$ |
| 1 erg | = | $6.242 \times 10^2 \text{ GeV}$ |
| 1 J | = | $6.242 \times 10^9 \text{ GeV}$ |
| 1 K | = | $8.618 \times 10^{-14} \text{ GeV}^{-1} / k_B$ |
| Astronomical unit | AU | $1.495\ 978\ 706\ 60(20) \times 10^{11} \text{ m}$ |
| parsec | pc | $3.085\ 677\ 580\ 7(4) \times 10^{16} \text{ m}$ |
| sidereal year | yr | $3.155\ 815 \times 10^7 \text{ s}$ |
| erg | erg | 10^{-7} J |
| gauss | gauss | 10^{-4} T |

A.4 Particle physics

| | | |
|--------------------------------|----------------------|--|
| Electron charge | e | $-1.602\ 176\ 487(40) \times 10^{-19} \text{ C}$ |
| Electron mass | m_e | $0.510\ 998\ 910(13) \text{ MeV}/c^2$ $= 9.109\ 389 \times 10^{-31} \text{ kg}$ |
| Proton mass | m_p | $938.271\ 013(23) \text{ MeV}/c^2$ $1.672\ 622 \times 10^{-27} \text{ kg}$ |
| Proton to electron mass ratio | $\mu = m_p/m_e$ | $1836.152\ 672\ 47(80)$ |
| Proton–neutron mass difference | $Q_{np} = m_n - m_p$ | $1.293\ 318(9) \text{ MeV}/c^2$ |
| Neutron to proton mass ratio | m_n/m_p | $1.001\ 378\ 419\ 18(46)$ |
| Mass of a deuterium nuclei | m_d | $1875.612\ 793(47) \text{ MeV}/c^2$ |
| Neutron lifetime | τ_n | $885.7(0.8) \text{ s}$ |

A.5 Cosmological quantities

| | | |
|---------------------------------------|---|---|
| Hubble constant | H_0 | $100 h \text{ km} \cdot \text{s}^{-1} \text{Mpc}^{-1}$ |
| | h | 0.72 ± 0.05 |
| Hubble time | $t_{H_0} = H_0^{-1}$ | $9.7776 h^{-1} \times 10^9 \text{ yr}$ $3.09 h^{-1} \times 10^{17} \text{ s}$ |
| Hubble distance | $D_{H_0} = cH_0^{-1}$ | $2997.9 h^{-1} \text{ Mpc}$ |
| Critical density today | $\rho_{\text{crit}} = \frac{3H_0^2}{8\pi G_N}$ | $1.878 37(28) \times 10^{-29} h^2 \text{ g} \cdot \text{cm}^{-3}$ $8.0992 \times 10^{-47} h^2 \text{ GeV}^4$ $1.0540 \times 10^4 h^2 \text{ eV} \cdot \text{cm}^{-3}$ |
| CMB temperature | $T_{\gamma 0}$ | $2.725 \Theta_{2.7} \text{ K}$ $2.348 \times 10^{-4} \Theta_{2.7} \text{ eV}$ |
| CMB photon number density | $n_{\gamma 0}$ | $410.44 \Theta_{2.7}^3 \text{ cm}^{-3}$ |
| CMB photon energy density | $\rho_{\gamma 0}$ | $4.640 8 \times 10^{-34} \Theta_{2.7}^4 \text{ g} \cdot \text{cm}^{-3}$ $0.2604 \Theta_{2.7}^4 \text{ eV} \cdot \text{cm}^{-3}$ |
| | $\Omega_{\gamma 0} h^2$ | $2.469 7 \times 10^{-5} \Theta_{2.7}^4$ |
| Neutrinos temperature | $T_{\nu 0} = (4/11)^{1/3} T_{\gamma 0}$ | $1.9272 \Theta_{2.7} \text{ K}$ |
| Neutrinos number density (per family) | $n_{\nu 0} = (3/11)n_{\gamma 0}$ | $108.89 \Theta_{2.7}^3 \text{ cm}^{-3}$ |
| Neutrinos energy density (per family) | $\rho_{\nu 0}$ | $0.2271 \rho_{\gamma 0}$ |
| Entropy | $s_0 = 7.0394 n_{\gamma 0}$ | $2889.2 \Theta_{2.7}^3$ |
| Radiation energy density | ρ_{r0} | $7.8042 \times 10^{-34} \Theta_{2.7}^4 \text{ g} \cdot \text{cm}^{-3}$ |
| | $\Omega_{r0} h^2$ | $4.153 4 \times 10^{-5} \Theta_{2.7}^4$ |
| Redshift at equality | $1 + z_{\text{eq}}$ | $2.4 \times 10^4 \Omega_{m0} h^2$ |
| Temperature at equality | $T_{\text{eq}} = T_{\gamma 0}(1 + z_{\text{eq}})$ | $6.55 \times 10^4 \Omega_{m0} h^2 \text{ K}$ |
| Redshift at decoupling | $1 + z_{\text{dec}}$ | 1100 |
| CMB temperature at decoupling | T_{dec} | $0.255 \Theta_{2.7} \text{ eV}$ |

A.6 Electromagnetic spectrum

The cosmic microwave background actually represents the dominant component of all radiation present in the Universe. The full cosmic background consists of the (large) volume average of all electromagnetic radiation present in the Universe. More precisely, it is defined as the isotropic background observed once the quasi-isotropic fluxes of both our Solar System and the galaxy have been subtracted.

The spectral distribution in energy has been observed from the radio band to the X-ray band and, as previously stated, the microwave component largely dominates over all other contributions, containing roughly 93% of all the energy. This cosmological microwave background is described in Section 4.4.2.

The radio background contains the sum of all extragalactic radiosources and was historically the first contribution to have been observed, although it contributes very little to the total radiative energy content. The X-ray component on the other hand,

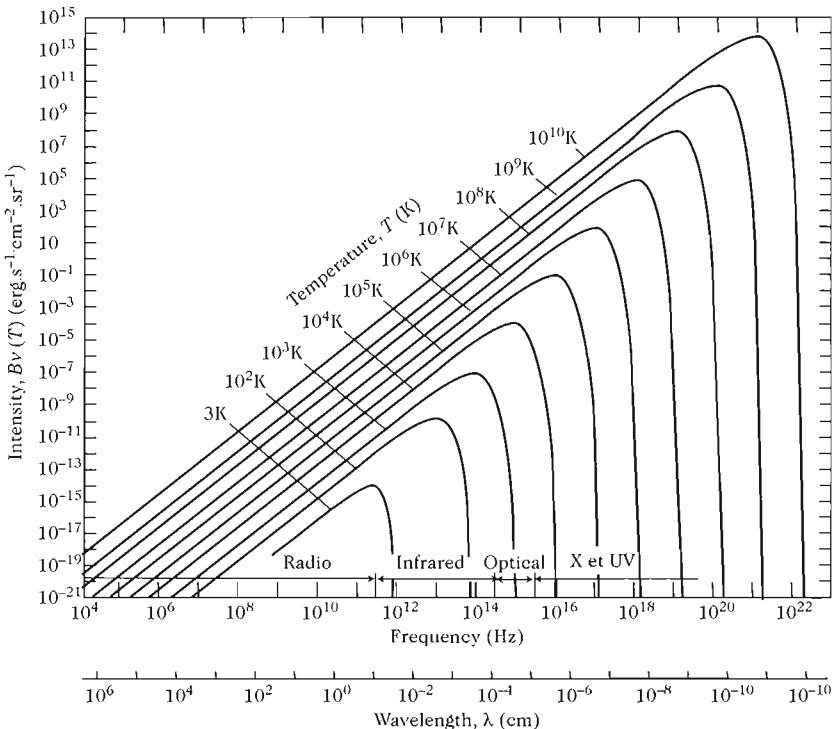


Fig. A.1 The luminosity $B_\nu(T)$ of blackbody radiation as a function of frequency ν (or, equivalently, of wavelength λ) for various temperatures T , including that of the cosmic microwave background (the lowest curve, marked with 3 K). Adapted from Ref. [3].

discovered by Giacconi and collaborators, is now very well measured. Its distribution peaks around 30 keV. The γ -ray background, in turn, is measured between 0.1 MeV and 1 GeV. Models for this component assume that it originates from the cores of ancient active galactic nuclei (AGN). The high energy tail of the γ distribution presumably also comes from similar sources. A fraction of the energy emitted in the far ultraviolet and the soft X-ray part of the spectrum is absorbed by intergalactic dust and re-emitted in the far infra-red. Hard X-rays, however, are not absorbed.

Infra-red and ultraviolet contributions are mainly due to galactic emission, integrated over all the galaxies. Optical and near infra-red radiation that is not absorbed by inter-galactic dust clouds dominates while diffused radiation contributes largely to the far infra-red.

The total energy density in electromagnetic radiation comes therefore mainly from the cosmic microwave background. The following table, as well as Fig. A.3, summarizes the contributions from the different sources of the extragalactic electromagnetic background.

| γ rays | Microwaves | | | | | | | |
|------------------------------|--------------------|--------------------|-------------------|-------------------|-------------------|-------------------|-------------------|---|
| X rays | UV | Visible | IR | mm | radio | | | |
| Hard | Soft | Far | Close | Close | Far | Sub | Radar | |
| $\leftarrow < 0.1\text{\AA}$ | 5\AA | 10nm | 200nm | 400nm | 700nm | 2.5\mu m | 25\mu m | 1mm |
| $> 10^9$ | 2×10^7 | 10^6 | 5×10^4 | 2.5×10^4 | 1.4×10^4 | 4000 | 400 | 10 |
| 3×10^{13} | 6×10^{12} | 3×10^{10} | 1.5×10^9 | 7.5×10^8 | 4×10^8 | 1.2×10^8 | 1.2×10^7 | 3×10^5 |
| 1.2×10^5 | 2400 | 120 | 6 | 3 | 1.7 | 0.5 | 0.05 | 0.001 |
| | | | | | | | | $10^{-5} \rightarrow \varepsilon/\text{eV}$ |

Fig. A.2 The electromagnetic radiation spectrum and its labelling (name of the radiation) according to its wavelength (λ), frequency (ν) or energy (ε).

Table A.1 Contribution of different wavelength bands to the electromagnetic background.

| Frequency band | Intensity ($\text{W.m}^{-2}.\text{sr}^{-1}$) | Contribution (%) |
|----------------|--|----------------------|
| Radio | 1.2×10^{-12} | 1.1×10^{-4} |
| CMB | 9.96×10^{-6} | 93 |
| Infra-red | $(4.6 \pm 0.6) \times 10^{-8}$ | 4.5 ± 0.5 |
| Optical | $(3 \pm 1) \times 10^{-8}$ | 3 ± 1 |
| X-ray | 2.7×10^{-10} | 3.5×10^{-2} |
| γ -ray | 3×10^{-11} | 2.5×10^{-3} |

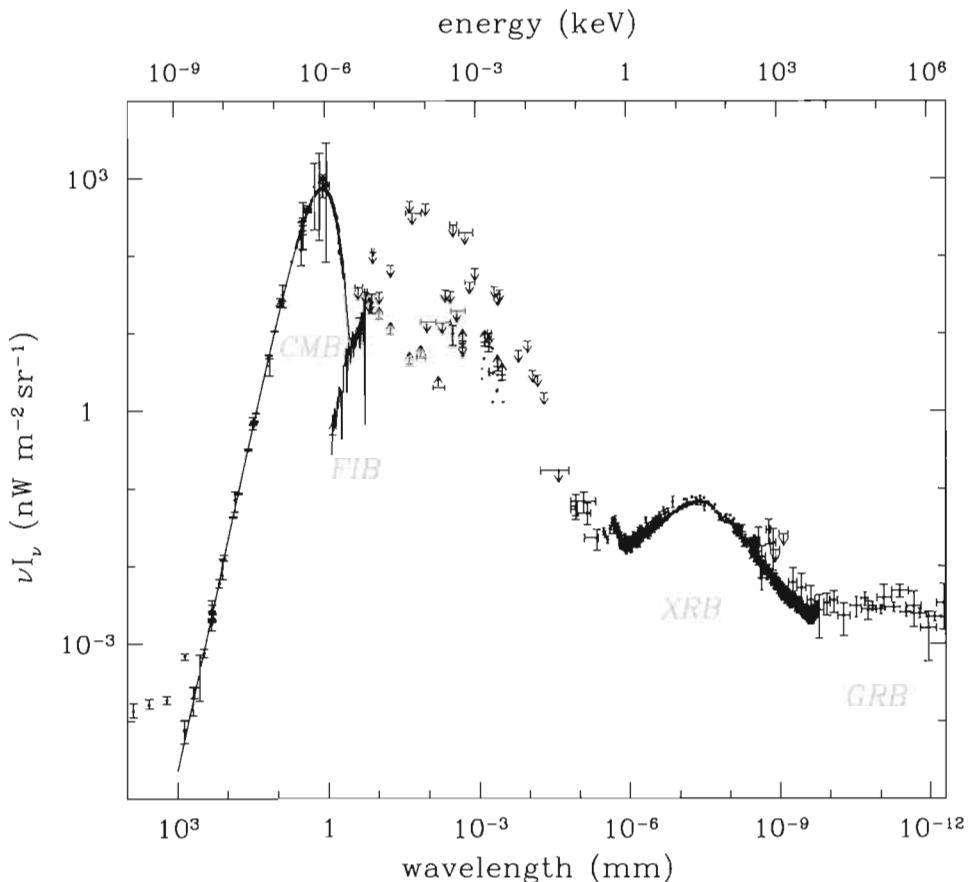


Fig. A.3 Spectral energy distribution for the radiation background, as compiled by Ref. [4]. Intensity is shown in logarithmic frequency bands so that the energy contained in each band can be directly compared. The most important contribution comes from the cosmic microwave background whose blackbody distribution is easily recognized (in the centimeter regime). Infra-red backgrounds are difficult to determine, mainly because of galactic contamination.

References

- [1] Recommended values of the fundamental constants (CODATA):
[\[http://physics.nist.gov/cuu/Constants/\]](http://physics.nist.gov/cuu/Constants/).
- [2] Particle Data Group: [\[http://pdg.lbl.gov/pdg.html\]](http://pdg.lbl.gov/pdg.html).
- [3] K. R. LANG, *Astrophysical formulae*, Springer-Verlag, 1980.
- [4] M. HALPERN and D. SCOTT, ‘Future microwave background experiments’, in *Microwave foregrounds*, A. de Oliveira-Costa and M. Tegmark (eds.), ASP, San Francisco, 1999.

Appendix B

Special functions

This appendix gathers the definitions and general properties of the special functions appearing in the core of the book. More details can be found in dedicated monographs, such as those in Refs. [1–3].

B.1 Euler functions

The Euler function of the second kind, also frequently called the Euler Γ function, is defined as

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt, \quad (\text{B.1})$$

for $\Re(z) > 0$. The Γ function is analytic with simple poles in $z = -p$, with $p \in \mathbb{N}$. For $n \in \mathbb{N}$, one recovers the well-known factorials,

$$\Gamma(n) = (n-1)! \quad (\text{B.2})$$

$$\Gamma\left(n + \frac{1}{2}\right) = \frac{\sqrt{\pi}}{2^n} (2n-1)!! \quad (\text{B.3})$$

$$\Gamma\left(-n + \frac{1}{2}\right) = (-1)^n \frac{2^n \sqrt{\pi}}{(2n-1)!!}. \quad (\text{B.4})$$

For large z , Γ behaves asymptotically as

$$\ln \Gamma(z) = \left(z - \frac{1}{2}\right) \ln z - z + \frac{1}{2} \ln 2\pi + \mathcal{O}(z^{-1}), \quad (\text{B.5})$$

which is simply the Weierstrass expansion formula.

The Euler function of first order, also called the Beta function, is defined by

$$B(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt. \quad (\text{B.6})$$

The Beta function is related to the Γ function through

$$B(x, y) = B(y, x) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \quad (\text{B.7})$$

B.2 Spherical harmonics

B.2.1 Definition

The spherical harmonics Y_ℓ^m are functions defined on the sphere. They are the eigenfunctions of the 2-dimensional spherical Laplacian. In other words, they satisfy

$$\left(\frac{\partial^2}{\partial\theta^2} + \cot\theta \frac{\partial}{\partial\theta} + \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial\varphi^2} \right) Y_\ell^m = -\ell(\ell+1)Y_\ell^m, \quad (\text{B.8})$$

where θ and φ are the angular coordinates on the surface of the sphere. The Y_ℓ^m form a complete set of functions, so they can be used as a basis for the functions on the sphere.

B.2.2 Expressions

Spherical harmonics are related to Legendre polynomials P_ℓ^m through – (5.2.1) in Ref. [1] –

$$Y_{\ell m}(\theta, \varphi) = \sqrt{\frac{2\ell+1}{4\pi}} \frac{(\ell-m)!}{(\ell+m)!} P_\ell^m(\cos\theta) e^{im\varphi}. \quad (\text{B.9})$$

They can also be expressed in terms of the Gegenbauer polynomials according to – (5.2.6.39c) of Ref. [1] –

$$Y_{\ell m}(\theta, \varphi) = \zeta_m e^{im\varphi} \sqrt{\frac{2\ell+1}{4\pi}} \sqrt{\frac{(\ell-|m|)!}{(\ell+|m|)!}} (2|m|-1)!! (\sin\theta)^{|m|} C_{\ell-|m|}^{|m|+1/2}(\cos\theta), \quad (\text{B.10})$$

ζ_m being defined as

$$\zeta_m = \begin{cases} (-1)^m & m > 0, \\ 1 & m \leq 0. \end{cases} \quad (\text{B.11})$$

The Gegenbauer polynomials C_n^α are solutions of the second-order differential equation

$$(1-x^2)y'' - (2\alpha+1)y' + n(n+2\alpha)y = 0, \quad (\text{B.12})$$

and satisfy the normalization relation – (7.313) of Ref. [2] –

$$\int_{-1}^1 (1-x^2)^{\alpha-1/2} [C_n^\alpha(x)]^2 dx = \frac{\pi 2^{1-2\alpha} \Gamma(2\alpha+n)}{n!(n+\alpha) [\Gamma(\alpha)]^2}, \quad (\text{B.13})$$

if $\Re e(\alpha) > -\frac{1}{2}$.

B.2.3 Properties

The complex conjugate of any given spherical harmonics is given by – (5.4.1) of Ref. [1] –

$$Y_{\ell m}^*(\theta, \varphi) = (-1)^m Y_{\ell -m}(\theta, \varphi) = Y_{\ell m}(\theta, -\varphi). \quad (\text{B.14})$$

Spherical harmonics are normalized according to – (5.6.1) of Ref. [1] –

$$\int_0^{2\pi} d\varphi \int_0^\pi \sin \theta d\theta Y_{\ell m}^*(\theta, \varphi) Y_{\ell' m'}(\theta, \varphi) = \delta_{\ell\ell'} \delta_{mm'}, \quad (\text{B.15})$$

leading to the closure relation – (5.2.2) of Ref. [1] –

$$\sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} Y_{\ell m}(\theta, \varphi) Y_{\ell m}^*(\theta', \varphi') = \delta^{(1)}(\cos \theta - \cos \theta') \delta^{(1)}(\varphi - \varphi'), \quad (\text{B.16})$$

as well as to the addition property – (5.17.2.9) of Ref. [1] –

$$\sum_{m=-\ell}^{\ell} Y_{\ell m}(\theta, \varphi) Y_{\ell m}^*(\theta', \varphi') = \frac{2\ell + 1}{4\pi} P_\ell(\cos \alpha), \quad (\text{B.17})$$

where α is the angle between the two directions (θ, φ) and (θ', φ') , and P_ℓ is the Legendre polynomial of order ℓ .

Any real function on the sphere is thus expanded on the spherical harmonics basis according to

$$f(\theta, \varphi) = \sum_{\ell, m} f_{\ell m} Y_{\ell m}(\theta, \varphi), \quad f_{\ell m} = (-1)^m f_{\ell, -m}. \quad (\text{B.18})$$

B.2.4 Fourier transform

The Fourier transform of spherical harmonics is given by – (5.9.2.6) of Ref. [1] –

$$\int_0^\infty r^2 dr \int_0^{2\pi} d\varphi \int_0^\pi \sin \theta d\theta \frac{e^{ik \cdot r}}{(2\pi)^{3/2}} j_\ell(k'r) Y_{\ell m}(\theta, \varphi) = \sqrt{\frac{2}{\pi}} i^\ell \frac{\delta^{(1)}(k' - k)}{k^2} Y_{\ell m}(\theta_k, \varphi_k), \quad (\text{B.19})$$

where j_ℓ is a spherical Bessel function (see below). From (B.19), one deduces the expansion of the exponential as – (5.17.3.14) of Ref. [1] –

$$e^{ik \cdot r} = \sum_{\ell=0}^{\infty} (2\ell + 1) i^\ell j_\ell(kr) P_\ell(\cos \theta_{k, r}), \quad (\text{B.20})$$

or, in other words,

$$e^{ip \cdot x} = 4\pi \sum_{\ell, m} i^\ell j_\ell(pr) Y_{\ell m}^*(\theta_p, \varphi_p) Y_{\ell m}(\theta_x, \varphi_x), \quad (\text{B.21})$$

where (θ_p, φ_p) and (θ_x, φ_x) represent the coordinates of the unit vectors parallel to p and x , respectively, and – (5.17.4.18) of Ref. [1] –

$$\delta^{(3)}(r_1 - r_2) = \frac{\delta^{(1)}(r_1 - r_2)}{r_1^2} \sum_{\ell=0}^{\infty} \frac{2\ell + 1}{4\pi} P_\ell(\cos \theta_{12}). \quad (\text{B.22})$$

B.2.5 Useful integrals

Integrating three spherical harmonics requires the so-called $3j$ -Wigner symbols,

$$\int d^2\gamma Y_{\ell_1 m_1}(\gamma) Y_{\ell_2 m_2}(\gamma) Y_{\ell_3 m_3}(\gamma) = \sqrt{\frac{(2\ell_1 + 1)(2\ell_2 + 1)(2\ell_3 + 1)}{4\pi}} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix}. \quad (\text{B.23})$$

These $3j$ -Wigner symbols,

$$\begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix},$$

are related to the Clebsch–Gordan coefficients (see Ref. [1] for their properties) and are non-zero provided

$$|\ell_1 - \ell_2| \leq \ell_3 \leq \ell_1 + \ell_2, \quad m_1 + m_2 + m_3 = 0. \quad (\text{B.24})$$

Furthermore, they satisfy the property

$$\sum_{m_1, m_2} \begin{pmatrix} \ell_1 & \ell_2 & \ell \\ m_1 & m_2 & m \end{pmatrix} \begin{pmatrix} \ell_1 & \ell_2 & \ell' \\ m_1 & m_2 & m' \end{pmatrix} = \frac{\delta_{\ell\ell'}\delta_{mm'}}{2\ell + 1}. \quad (\text{B.25})$$

With the above properties at hand, the integrals we have encountered in Chapter 6 are easily calculated. We have

$$B_{\ell m}^{\pm 1} = \int d^2\gamma Y_{\ell m}^*(\gamma) e^{i\gamma \cdot k} \sin \theta e^{\pm i\varphi}, \quad (\text{B.26})$$

where γ is the unit vector pointing in the direction (θ, φ) . Relation (8.733) in Ref. [2] tells us that

$$B_{\ell m}^{\pm 1} = (-1)^{\ell+1} \delta_{m \pm 1} \sqrt{4\pi(2\ell+1)} \sqrt{\frac{\ell(\ell+1)}{2}} \frac{j_\ell(k)}{k}. \quad (\text{B.27})$$

Analogously, the relation

$$A_{\ell m}^\lambda = \int d^2\gamma Y_{\ell m}^*(\gamma) e^{i\gamma \cdot k} \sin^2 \theta (\delta_\times^\lambda \sin 2\varphi + \delta_+^\lambda \cos 2\varphi) \quad (\text{B.28})$$

reduces to

$$A_{\ell m}^\lambda = \pi \sqrt{\frac{2\ell+1}{4\pi}} \sqrt{\frac{(\ell+2)!}{(\ell-2)!}} \alpha_{\ell m}^\lambda \sum_n i^n (2n+1) j_\ell(k) (c_{-2} \delta_{\ell-2}^n - 2c_0 \delta_\ell^n + c_{+2} \delta_{\ell+2}^n), \quad (\text{B.29})$$

with coefficients

$$\alpha_{\ell m}^\lambda = \delta_m^2 (\delta_+^\lambda - i\delta_\times^\lambda) + \delta_m^{-2} (\delta_+^\lambda + i\delta_\times^\lambda),$$

and

$$c_n^{-1} = \frac{1}{2}(2\ell-1+n)(2\ell+1+n)(2\ell+3+n).$$

B.3 Bessel functions

B.3.1 Definition

Bessel functions, generically denoted by $Z_\nu(z)$, are solutions of the differential equation

$$\frac{d^2Z_\nu}{dz^2} + \frac{1}{z} \frac{dZ_\nu}{dz} + \left(1 - \frac{\nu^2}{z^2}\right) Z_\nu = 0. \quad (\text{B.30})$$

It is common to distinguish between Bessel functions of the first (called J_ν) and the second (N_ν) kind, the latter being also often called the Neumann functions. They are defined, respectively, through

$$J_\nu(z) = \frac{z^\nu}{2^\nu} \sum_{k=0}^{\infty} (-1)^k \frac{z^{2k}}{2^{2k} k! \Gamma(\nu + k + 1)} \quad (\text{B.31})$$

if $|\arg z| > \pi$, and

$$N_\nu = \frac{1}{\sin \nu \pi} [\cos \nu \pi J_\nu(z) - J_{-\nu}(z)] \quad (\text{B.32})$$

if ν is not an integer.

Integer-order ($\nu = n \in \mathbb{Z}$) Bessel functions of the first and second kind enjoy the symmetry

$$J_{-n}(z) = (-1)^n J_n(z), \quad N_{-n}(z) = (-1)^n N_n(z). \quad (\text{B.33})$$

The two functions J_ν and N_ν are two linearly independent solutions of (B.30).

It is sometimes useful to introduce the Bessel function of the third kind, also called Hankel functions, defined by

$$H_\nu^{(1)}(z) = J_\nu(z) + iN_\nu(z), \quad H_\nu^{(2)}(z) = J_\nu(z) - iN_\nu(z). \quad (\text{B.34})$$

Generic Bessel functions satisfy the recursion relations

$$zZ_{\nu-1} + zZ_{\nu+1} = 2\nu Z_\nu, \quad (\text{B.35})$$

$$Z_{\nu-1} - Z_{\nu+1} = 2Z'_\nu. \quad (\text{B.36})$$

B.3.2 Asymptotic properties

One can expand the functions J_ν as series according to

$$J_\nu(z) = \left(\frac{z}{2}\right)^\nu \sum_{k=0}^{\infty} (-1)^k \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{z}{2}\right)^{2k}. \quad (\text{B.37})$$

From this, one gets their behaviour at the origin, i.e. for small values of the argument (around $z \sim 0$),

$$J_\nu(z) \sim \frac{1}{\Gamma(\nu + 1)} \left(\frac{z}{2}\right)^\nu, \quad N_\nu(z) \sim \frac{1}{\Gamma(-\nu + 1) \sin \nu \pi} \left(\frac{z}{2}\right)^{-\nu}. \quad (\text{B.38})$$

At infinity, on the other hand, Bessel functions have the following asymptotic behaviours

$$J_{\pm\nu} \sim -\sqrt{\frac{2}{\pi z}} \cos\left(z \pm \frac{\pi}{2}\nu - \frac{\pi}{4}\right), \quad N_{\pm\nu} \sim -\sqrt{\frac{2}{\pi z}} \sin\left(z \pm \frac{\pi}{2}\nu - \frac{\pi}{4}\right), \quad (\text{B.39})$$

and hence those of the Hankel functions are

$$H_{\nu}^{(1)}(z) \sim \sqrt{\frac{2}{\pi z}} e^{i(z - \frac{\pi}{2}\nu - \frac{\pi}{4})}, \quad H_{\nu}^{(2)}(z) \sim \sqrt{\frac{2}{\pi z}} e^{-i(z - \frac{\pi}{2}\nu - \frac{\pi}{4})}. \quad (\text{B.40})$$

B.3.3 Special cases

Some Bessel functions can be expressed in terms of known simple functions for some values of the index ν . For instance, one has

$$J_{\frac{1}{2}} = \sqrt{\frac{2}{\pi z}} \sin z, \quad N_{\frac{1}{2}} = \sqrt{\frac{2}{\pi z}} \cos z, \quad (\text{B.41})$$

$$H_{\frac{1}{2}}^{(1)} = \sqrt{\frac{2}{\pi z}} \frac{e^{iz}}{i}, \quad H_{-\frac{1}{2}}^{(2)} = \sqrt{\frac{2}{\pi z}} \frac{e^{-iz}}{-i}, \quad (\text{B.42})$$

and many other similar simple forms for a half-integer ν .

It is of interest to write down at this point the solution to some differential equations frequently met in cosmology in terms of Bessel functions. For instance, we have

$$u'' + \left[(\beta\gamma z^{\gamma-1})^2 - \left(\frac{\nu\gamma}{z}\right)^2 \right] u = 0 \implies u = z Z_{\nu}(\beta z^{\gamma}), \quad (\text{B.43})$$

$$u'' + \frac{1-2\alpha}{z} u' + \left[(\beta\gamma z^{\gamma-1})^2 + \frac{\alpha^2 - \nu^2\gamma^2}{z^2} \right] u = 0 \implies u = z^{\alpha} Z_{\nu}(\beta z^{\gamma}), \quad (\text{B.44})$$

and it should be emphasized that a particularly important case in cosmology, as the reader undoubtedly remarked, is that for which $\alpha = -\frac{1}{2}$ and $\gamma = 1$.

B.3.4 Spherical Bessel functions

Spherical Bessel functions are defined through the relation

$$j_{\nu}(z) \equiv \sqrt{\frac{\pi}{2z}} J_{\nu+\frac{1}{2}}(z), \quad (\text{B.45})$$

and satisfy the recursion relations

$$\begin{aligned} j_{\ell} &= \frac{z}{2\ell+1} (j_{\ell-1} + j_{\ell+1}), \\ j'_{\ell} &= \frac{1}{2\ell+1} [\ell j_{\ell-1} - (\ell+1) j_{\ell+1}], \\ j'_{\ell} &= j_{\ell-1} - \frac{\ell+1}{z} j_{\ell}, \end{aligned} \quad (\text{B.46})$$

from which one deduces

$$\frac{j_{\ell}}{z^2} = \frac{j_{\ell-2}}{(2\ell-1)(2\ell+1)} + 2 \frac{j_{\ell}}{(2\ell-1)(2\ell+3)} + \frac{j_{\ell+2}}{(2\ell+1)(2\ell+3)}. \quad (\text{B.47})$$

It can be checked from (B.44) that they satisfy the second-order differential equation

$$j_\ell''(x) + \frac{2}{x} j_\ell'(x) + \left[1 - \frac{\ell(\ell+1)}{x^2} \right] j_\ell(x) = 0. \quad (\text{B.48})$$

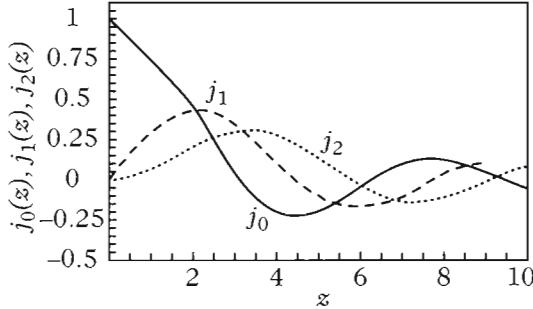


Fig. B.1 Spherical Bessel functions for $\ell = 0, 1, 2$.

B.3.5 Some useful integrals

The Bessel functions satisfy

$$\int_0^\infty J_\nu(t) J_\mu(t) t^{-\lambda} dt = \frac{2^{-\lambda} \Gamma(\lambda) \Gamma\left(\frac{\nu+\mu-\lambda+1}{2}\right)}{\Gamma\left(\frac{-\nu+\mu+\lambda+1}{2}\right) \Gamma\left(\frac{\nu+\mu+\lambda+1}{2}\right) \Gamma\left(\frac{\nu-\mu+\lambda+1}{2}\right)}, \quad (\text{B.49})$$

if $\Re(\nu + \mu + 1) > \Re(\lambda) > 0$. This means that

$$\int_0^\infty j_\ell(t) j_\ell(t) t^{-k} dt = \frac{\pi}{2} \frac{\Gamma(k+1) \Gamma\left(\frac{2\ell+1-k}{2}\right)}{2^{k+1} \Gamma^2\left(\frac{k+2}{2}\right) \Gamma\left(\frac{2\ell+k+3}{2}\right)}, \quad (\text{B.50})$$

with $-1 < k < 2\ell + 1$. This also leads to

$$\int_0^\infty j_\ell(pr) j_{\ell'}(p'r) r^2 dr = \frac{\pi}{2p^2} \delta(p - p'), \quad (\text{B.51})$$

defining the normalization of the spherical Bessel function.

B.4 Legendre polynomials

B.4.1 Associated Legendre polynomials

An associated Legendre polynomial, $P_\nu^\mu(z)$, is a solution of the differential equation

$$(1 - z^2) \frac{d^2u}{dz^2} - 2z \frac{du}{dz} + \left[\nu(\nu + 1) - \frac{\mu^2}{1 - z^2} \right] u = 0, \quad (\text{B.52})$$

μ and ν being two complex numbers. These polynomials satisfy the recursion relation – (8.733) of Ref. [2] –

$$P_{\nu-1}^\mu - P_{\nu+1}^\mu = (2\nu + 1)\sqrt{1 - x^2}P_\nu^{\mu-1}. \quad (\text{B.53})$$

B.4.2 Legendre polynomials

Legendre polynomials are in fact a special case of associated Legendre polynomials for which $\mu = 0$ and $\nu = n \in \mathbb{N}$. These polynomials form an orthogonal basis for the functions defined on the segment $[-1, 1]$, satisfying the orthogonality relation

$$\int_{-1}^1 \left(n + \frac{1}{2} \right)^{\frac{1}{2}} P_n(x) \left(m + \frac{1}{2} \right)^{\frac{1}{2}} P_m(x) dx = \delta_{mn}. \quad (\text{B.54})$$

In particular, one has

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{1}{2}(3x^2 - 1), \quad P_3(x) = \frac{x}{2}(5x^2 - 3).$$

Any given function of the variable $\cos \varphi$ can be expanded as

$$f(\cos \varphi) = \sum_n \frac{2n+1}{2} a_n P_n(\cos \varphi). \quad (\text{B.55})$$

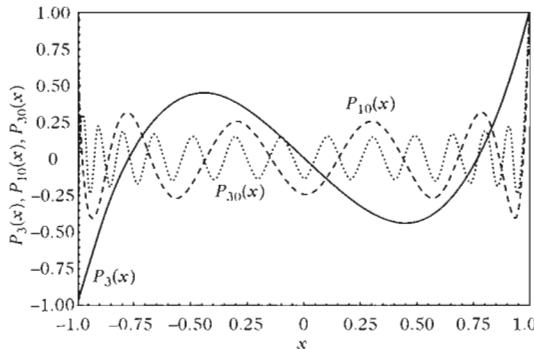


Fig. B.2 Legendre polynomials for $\ell = 3, 10, 30$.

B.5 Fourier transform and the eigenmodes of the Laplacian

B.5.1 Fourier transform

In a D -dimensional maximally symmetric space, the Helmholtz equation is an eigenvalue equation of the Laplacian

$$\Delta Q_{\mathbf{k}} = -(k^2 - K)Q_{\mathbf{k}}, \quad (\text{B.56})$$

where the wavevector \mathbf{k} labels the modes and the sign of K depends on the spatial curvature; Δ is the Laplacian defined from the covariant derivative D_i associated to the spatial metric γ_{ij} . We also impose the normalization condition

$$\int Q_{\mathbf{k}}(\mathbf{x})Q_{\mathbf{k}'}^*(\mathbf{x})\sqrt{\gamma}d^Dx = (2\pi)^D\delta^{(D)}(\mathbf{k} - \mathbf{k}'). \quad (\text{B.57})$$

Any function can be expanded in Fourier modes on the basis of the functions $Q_{\mathbf{k}}(\mathbf{x})$ according to

$$f(\mathbf{x}) = \int \frac{d^D k}{(2\pi)^{D/2}} \hat{f}(\mathbf{k}) Q_{\mathbf{k}}(\mathbf{x}), \quad (\text{B.58})$$

where $\hat{f}(\mathbf{k})$ are the Fourier coefficients of the function $f(\mathbf{x})$. This relation is inverted using (B.57), yielding

$$\hat{f}(\mathbf{k}) = \int \sqrt{\gamma} \frac{d^D x}{(2\pi)^{D/2}} f(\mathbf{x}) Q_{\mathbf{k}}^*(\mathbf{x}). \quad (\text{B.59})$$

B.5.2 Power spectrum

If $f(\mathbf{x})$ is a given homogeneous and isotropic stochastic field, one can define its correlation function by

$$\xi_f(r) = \langle f(\mathbf{x})f(\mathbf{x} + \mathbf{r}) \rangle. \quad (\text{B.60})$$

Its power spectrum is defined in Fourier space as

$$\langle \hat{f}(\mathbf{k}) \hat{f}^*(\mathbf{k}') \rangle = \delta^{(D)}(\mathbf{k} - \mathbf{k}') P_f(k). \quad (\text{B.61})$$

It turns out to be handy to define a spectrum with different dimensions, namely

$$\mathcal{P}_f(k) = \frac{2S^{[D-2]}}{(2\pi)^D} k^D P_f(k), \quad (\text{B.62})$$

where $S^{[D-1]}$ is the surface of the $(D-1)$ -dimensional sphere of unit radius,

$$S^{[D-1]} = \frac{2\pi^{D/2}}{\Gamma(D/2)}, \quad (\text{B.63})$$

where we recover the particular value for the surface of the 1-sphere, i.e. a circle, $S^{[1]} = 2\pi$, and that of the 2-sphere, for which $S^{[2]} = 4\pi$.

The correlation function stems from the power spectrum through

$$\xi_f(r) = \frac{2S^{[D-2]}}{(2\pi)^D} \int P_f(k) k^D \frac{\sin kr}{kr} \frac{dk}{k} = \int \mathcal{P}_f(k) \frac{\sin kr}{kr} \frac{dk}{k}, \quad (\text{B.64})$$

after integrating out over the $(D-1)$ angular directions.

B.5.3 Cartesian coordinates

To get the explicit form of the functions $Q_{\mathbf{k}}$, let us restrict attention to the case $D = 3$. Depending on the curvature sign, the wave numbers take the values

$$K < 0 : \quad k \in [0, \infty[\quad \text{and} \quad ik \in [0, \sqrt{-K}], \quad (\text{B.65})$$

$$K = 0 : \quad k \in [0, \infty[, \quad (\text{B.66})$$

$$K > 0 : \quad k = (\nu + 1)\sqrt{K}, \quad \nu \in \mathbb{N}. \quad (\text{B.67})$$

In Cartesian coordinates, only the functions for $K = 0$ take a simple form, namely

$$Q_{\mathbf{k}}(\mathbf{x}) = \exp(i\mathbf{k} \cdot \mathbf{x}), \quad (\text{B.68})$$

and Eqs. (B.58)–(B.59) take the usual form of the Fourier transform. Note also that, still assuming flat space $K = 0$, we get the following useful expression

$$\delta^{(D)}(\mathbf{x} - \mathbf{x}') = \frac{1}{(2\pi)^D} \int d^D \mathbf{k} e^{i\mathbf{k} \cdot (\mathbf{x} - \mathbf{x}')}, \quad (\text{B.69})$$

for the Dirac distribution $\delta^{(D)}$ in D dimensions.

B.5.4 Spherical coordinates

In a spherical coordinate system (χ, θ, φ) , the eigenfunctions can be expanded as

$$Q_{k\ell m} = (2\pi)^{3/2} R_{k\ell}(\chi) Y_{\ell m}(\theta, \varphi), \quad (\text{B.70})$$

with the normalization taking the form

$$\int Q_{k\ell m}(\chi, \theta, \varphi) Q_{k' \ell' m'}^*(\chi, \theta, \varphi) f_K^2(\chi) d\chi d\Omega = \frac{(2\pi)^3}{k^2} \delta^{(3)}(k - k') \delta_{\ell\ell'} \delta_{mm'}. \quad (\text{B.71})$$

With the decomposition (B.70), the radial part of the Helmholtz equation becomes

$$\frac{1}{f_K^2(\chi)} \frac{d}{d\chi} \left[f_K^2(\chi) \frac{d}{d\chi} R_{k\ell} \right] B + \left[(k^2 - K) - \frac{\ell(\ell+1)}{f_K^2(\chi)} \right] R_{k\ell} = 0. \quad (\text{B.72})$$

This second-order differential equation has two independent solutions, only one of which is regular at $\chi = 0$, so that the radial function is fully determined once the normalization is chosen. Solving this equation yields the following three solutions [4,5]

$$R_{k\ell}(\chi) = \sqrt{\frac{N_{k\ell}}{kf_K(\chi)}} P_{-1/2+\ell}^{-1/2-\ell} [\cosh(\sqrt{-K}\chi)] \quad K < 0 \quad (\text{B.73})$$

$$= \sqrt{\frac{2}{\pi}} j_\ell(k\chi) \quad K = 0 \quad (\text{B.74})$$

$$= \sqrt{\frac{M_{k\ell}}{kf_K(\chi)}} P_{-1/2+\ell}^{-1/2-\ell} [\cos(\sqrt{K}\chi)] \quad K > 0. \quad (\text{B.75})$$

The numerical normalization coefficients are given by

$$N_{k\ell} = \prod_{n=0}^{\ell} (\omega^2 + n^2), \quad M_{k\ell} = \begin{cases} \prod_{n=0}^{\ell} [(\nu + 1)^2 - n^2] & \text{for } \ell \leq \nu, \\ 0 & \text{for } \ell > \nu, \end{cases} \quad (\text{B.76})$$

with

$$\omega = k\sqrt{-K}, \quad \nu = \frac{k}{\sqrt{K}} - 1. \quad (\text{B.77})$$

Note that in the case $K < 0$, the normalization condition only holds for the modes for which k is real [6], which correspond to wavelengths smaller than the curvature radius of space. For the supercurvature modes, i.e. $ik \in [0, \sqrt{-K}]$, the radial function can be obtained by analytic continuation.

Any function can then be expanded into spherical Fourier modes as

$$f(\chi, \theta, \varphi) = \frac{1}{(2\pi)^{3/2}} \sum_{\ell m} \left\{ \frac{\int_0^\infty k^2 dk}{K^{3/2} \sum_{\nu=2}^\infty (\nu + 1)^2} \right\} f_{\ell m}(k) Q_{k\ell m}(\chi, \theta, \varphi), \quad (\text{B.78})$$

the Fourier coefficients being given by

$$f_{\ell m}(k) = \int f(\chi, \theta, \varphi) Q_{k\ell m}^*(\chi, \theta, \varphi) f_K^2(\chi) \frac{d\chi d\Omega}{(2\pi)^{3/2}}. \quad (\text{B.79})$$

Using Eqs. (B.15) and (B.21), it can be shown that these Fourier coefficients are related by

$$f_{\ell m}(k) = (-i)^\ell \int \hat{f}(\mathbf{k}) Y_{\ell m}(\hat{\mathbf{k}}) d\Omega_{\mathbf{k}}. \quad (\text{B.80})$$

References

- [1] D.A. VARSHALOVICH, A.N. MOSKALEV, and V.K. KHERSONSKII, *Quantum theory of angular momentum*, World Scientific, Singapore, 1988.
- [2] I.S. GRADSHTEYN and I.M. RYZHIK, *Table of integrals, series and products*, Academic Press, 1980.
- [3] M. ABRAMOWITZ and I.A. STEGUN, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Dover Publications, 1970.
- [4] L.F. ABOTT and R.K SCHAEFFER, ‘A general, gauge invariant analysis of the cosmic microwave background’, *Astrophys. J.* **308**, 546, 1986, [Appendix].
- [5] A. RIAZUELO *et al.*, ‘Simulating cosmic microwave background maps in multiconnected space’, *Phys. Rev. D* **69**, 103514, 2004. [Appendix A].
- [6] D.H. LYTH and A. WOSZCZYNA, ‘Large scale perturbations in the open Universe’, *Phys. Rev. D* **52**, 3338, 1995.



Appendix C

Useful cosmological quantities

C.1 Background space

C.1.1 Geometry

The Friedmann–Lemaître metric is given, in conformal time, by

$$\begin{aligned} g_{00} &= -a^2, \\ g_{0i} &= 0, \\ g_{ij} &= a^2 \gamma_{ij}, \end{aligned} \tag{C.1}$$

where the spatial metric γ_{ij} can be written as

$$\gamma_{ij} = \frac{\delta_{ij}}{\left(1 + \frac{1}{4}K\delta_{mn}x^m x^n\right)^2}, \tag{C.2}$$

and we have the inverse metric

$$\begin{aligned} g^{00} &= -a^{-2}, \\ g^{0i} &= 0, \\ g^{ij} &= a^{-2}\gamma^{ij}. \end{aligned} \tag{C.3}$$

Note that the spatial metric can also take the form

$$\gamma_{ij}dx^i dx^j = \frac{dr^2}{1-Kr^2} + r^2 d\Omega^2,$$

in spherical coordinates or

$$\gamma_{ij}dx^i dx^j = d\chi^2 + f_K^2(\chi)d\Omega^2, \tag{C.4}$$

with

$$f_K(\chi) = \begin{cases} \frac{1}{\sqrt{K}} \sin(\sqrt{K}\chi), \\ \chi, \\ \frac{1}{\sqrt{-K}} \sinh(\sqrt{-K}\chi), \end{cases} \tag{C.5}$$

respectively for K positive, null and negative.

The non-vanishing Christoffel symbols are

$$\begin{aligned}\Gamma_{00}^0 &= \mathcal{H}, \\ \Gamma_{ij}^0 &= \mathcal{H}\gamma_{ij}, \\ \Gamma_{0i}^j &= \mathcal{H}\delta_i^j, \\ \Gamma_{jk}^i &= \gamma_{jk}^i,\end{aligned}\tag{C.6}$$

where γ_{jk}^i are the Christoffel symbols derived from the spatial metric γ_{ij} .

The non-vanishing components of the Riemann tensor are given by

$$\begin{aligned}R^0_{i0j} &= \mathcal{H}'\gamma_{ij}, \\ R^i_{00j} &= \mathcal{H}'\delta_j^i, \\ R^i_{jml} &= (\mathcal{H}^2 + K)(\delta_m^i\gamma_{jl} - \delta_l^i\gamma_{jm}),\end{aligned}\tag{C.7}$$

leading to those of the Ricci tensor

$$\begin{aligned}R_{00} &= -3\mathcal{H}', \\ R_{ij} &= (2\mathcal{H}^2 + \mathcal{H}' + 2K)\gamma_{ij},\end{aligned}\tag{C.8}$$

and eventually the scalar curvature

$$Ra^2 = 6(\mathcal{H}^2 + \mathcal{H}' + K).\tag{C.9}$$

Using these tensors, one obtains the non-vanishing components of the Einstein tensor, namely

$$\begin{aligned}G_{00} &= 3(\mathcal{H}^2 + K), \\ G_{ij} &= -(\mathcal{H}^2 + 2\mathcal{H}' + K)\gamma_{ij}.\end{aligned}\tag{C.10}$$

The spatial curvature is given by

$${}^{(3)}R = 6K,\tag{C.11}$$

which is computed by contracting the Riemann, and then the Ricci tensors associated to the metric γ_{ij} , namely

$${}^{(3)}R^i_{jkl} = K(\delta_k^i\gamma_{jl} - \delta_l^i\gamma_{jk}), \quad {}^{(3)}R^{ij} = {}^{(3)}R^k_{ikj} = 2K\gamma_{ij},\tag{C.12}$$

respectively.

C.1.2 Matter

The stress-energy tensor for a perfect fluid of density ρ and pressure P is

$$T_{\mu\nu} = \rho u_\mu u_\nu + P(g_{\mu\nu} + u_\mu u_\nu),\tag{C.13}$$

with normalized 4-velocity $u_\mu u^\mu = -1$ so that one can take

$$u^0 = \frac{1}{a}, \quad u^i = 0, \quad u_0 = -a, \quad u_i = 0. \quad (\text{C.14})$$

The non-vanishing components of this tensor are then given by

$$\begin{aligned} T_{00} &= \rho a^2, \\ T_{ij} &= P a^2 \gamma_{ij}, \end{aligned} \quad (\text{C.15})$$

or equivalently,

$$\begin{aligned} T_0^0 &= -\rho, \\ T_j^i &= P \delta_j^i. \end{aligned} \quad (\text{C.16})$$

The conservation for a fluid takes the form

$$\rho' + 3\mathcal{H}(1+w)\rho = 0, \quad (\text{C.17})$$

where the equation of state w is defined by

$$P = w\rho. \quad (\text{C.18})$$

After some algebra, we obtain that

$$w' = -3\mathcal{H}(1+w)(c_s^2 - w), \quad (\text{C.19})$$

with the speed of sound c_s defined

$$c_s^2 = \frac{P'}{\rho'}. \quad (\text{C.20})$$

Using the Einstein equations, it turns out to be extremely useful to express both the pressure and the density as functions of the geometric quantities as

$$\kappa\rho a^2 = 3(\mathcal{H}^2 + K), \quad \kappa P a^2 = -(\mathcal{H}^2 + 2\mathcal{H}' + K), \quad \kappa(\rho+3P)a^2 = -6\mathcal{H}', \quad (\text{C.21})$$

from where we obtain the following formula after some calculations

$$\mathcal{H}' = -\frac{1}{2}(\mathcal{H}^2 + K)(1 + 3w). \quad (\text{C.22})$$

For a scalar field, we have

$$\begin{aligned} T_{00} &= \frac{1}{2}\varphi'^2 + a^2V(\varphi), \\ T_{ij} &= \left[\frac{1}{2}\varphi'^2 - a^2V(\varphi)\right]\gamma_{ij}. \end{aligned} \quad (\text{C.23})$$

C.2 Perturbed quantities

C.2.1 Geometry

The perturbed Friedmann–Lemaître metric in conformal time is given by

$$\begin{aligned}\delta g_{00} &= -2a^2 A, \\ \delta g_{0i} &= a^2 (D_i B + \bar{B}_i) \equiv a^2 B_i, \\ \delta g_{ij} &= a^2 [2C\gamma_{ij} + 2D_i D_j E + 2D_{(i} \bar{E}_{j)} + 2\bar{E}_{ij}] \equiv a^2 h_{ij},\end{aligned}\quad (\text{C.24})$$

where we have explicitly separated out the scalar, vector and tensor parts (SVT decomposition). Let us recall that D_i represents the covariant derivative with respect to the spatial metric γ_{ij} , and that latin indices (i, j, \dots) are raised and lowered by means of the background spatial metric γ_{ij} . The vectors \bar{B}_i and \bar{E}_i are divergenceless ($D_i \bar{B}^i = D_i \bar{E}^i = 0$), and so is the tensor \bar{E}_{ij} , which is also traceless ($D_i \bar{E}^{ij} = 0$ and $\bar{E}^i_i = 0$). At the same order in perturbations the inverse metric is

$$\begin{aligned}\delta g^{00} &= \frac{2A}{a^2}, \\ \delta g^{0i} &= \frac{1}{a^2} (D^i B + \bar{B}^i), \\ \delta g^{ij} &= -\frac{1}{a^2} [2C\gamma^{ij} + 2D^i D^j E + 2D^{(i} \bar{E}^{j)} + 2\bar{E}^{ij}],\end{aligned}\quad (\text{C.25})$$

from which one derives the perturbed Christoffel symbols as

$$\begin{aligned}\delta\Gamma_{00}^0 &= A', \\ \delta\Gamma_{00}^i &= D^i (A + B' + \mathcal{H}B) + \bar{B}^{i'} + \mathcal{H}\bar{B}^i, \\ \delta\Gamma_{0i}^0 &= D_i (A + \mathcal{H}B) + \mathcal{H}\bar{B}_i, \\ \delta\Gamma_{ij}^0 &= -2\mathcal{H}A\gamma_{ij} + \mathcal{H}h_{ij} + \frac{1}{2}h'_{ij} - D_{(i} D_{j)} B - D_{(i} \bar{B}_{j)}, \\ \delta\Gamma_{0j}^i &= \frac{1}{2}h_j^{i'} + \gamma^{il} D_{[j} \bar{B}_{l]}, \\ \delta\Gamma_{jk}^i &= -\mathcal{H} (D^i B + \bar{B}^i) \gamma_{jk} + D_{(j} h_{k)}^i - \frac{1}{2}D^i h_{jk}.\end{aligned}\quad (\text{C.26})$$

The perturbed Christoffel symbols allow us to compute the perturbed Ricci tensor. This is

$$\begin{aligned}\delta R_{00} &= 3\mathcal{H}A' + \Delta A + \Delta(B' + \mathcal{H}B) - \frac{1}{2}(h'' + \mathcal{H}h'), \\ \delta R_{0i} &= 2\mathcal{H}D_i A + (\mathcal{H}' + 2\mathcal{H}^2)B_i + D^k D_{[i} B_{k]} + \frac{1}{2}(D_k h_i^{ik} - D_i h'), \\ \delta R_{ij} &= -[2(2\mathcal{H}^2 + \mathcal{H}')A + \mathcal{H}A' + \mathcal{H}\Delta B] \gamma_{ij} - D_i D_j A - D_{(i} [B'_{j)} + 2\mathcal{H}B_{j)}] \\ &\quad + (2\mathcal{H}^2 + \mathcal{H}')h_{ij} + \mathcal{H}h'_{ij} + \frac{1}{2}\mathcal{H}h'\gamma_{ij} + \frac{1}{2}h''_{ij} \\ &\quad + D_{(i} D_k h_{j)}^k - \frac{1}{2}\Delta h_{ij} - \frac{1}{2}D_i D_j h,\end{aligned}\quad (\text{C.27})$$

leading, upon contraction, to the curvature perturbation

$$a^2 \delta R = h'' + 3\mathcal{H}h' - 4\Delta C - 2Kh - 2\Delta A - 12(\mathcal{H}' + \mathcal{H}^2)A - 6\mathcal{H}A' - 2\Delta(B' + 3\mathcal{H}B), \quad (\text{C.28})$$

where use has been made of the trace relation $h = 6C + 2\Delta E$, and of the formula $\Delta h - D^{(i}D^{j)}h_{ij} = 4\Delta C$.

The perturbed Einstein tensor is then given by

$$\begin{aligned} \delta G_{00} &= 6\mathcal{H}C' - 2\Delta C + 6K(A - C) - 2\mathcal{H}\Delta(B - E'), \\ \delta G_{0i} &= D_i [2\mathcal{H}A - 2C' + 2KE' - (2\mathcal{H}' + \mathcal{H}^2 + 3K)B] \\ &\quad + \frac{1}{2}\Delta(\bar{E}'_i - \bar{B}_i) + K\bar{E}'_i - (2\mathcal{H}' + \mathcal{H}^2 + 2K)\bar{B}_i, \\ \delta G_{ij} &= D_i D_j [E'' + 2\mathcal{H}E' - 2(2\mathcal{H}' + \mathcal{H}^2 + K)E - B' - 2\mathcal{H}B - C - A] \\ &\quad + \gamma_{ij} [-\Delta(E'' + 2\mathcal{H}E') + \Delta(B' + 2\mathcal{H}B) - 2C'' - 4\mathcal{H}C' + \Delta(C + A) \\ &\quad + 2(2\mathcal{H}' + \mathcal{H}^2)(A - C) + 2\mathcal{H}A'] \\ &\quad + D_{(i} \left\{ [\bar{E}'_{j)} - \bar{B}_{j)}]' + 2\mathcal{H}[\bar{E}'_{j)} - \bar{B}_{j)}] \right\} - 2(2\mathcal{H}' + \mathcal{H}^2 + K)\bar{E}_{j)} \\ &\quad + \bar{E}''_{ij} + 2\mathcal{H}\bar{E}'_{ij} - 2(2\mathcal{H}' + \mathcal{H}^2)\bar{E}_{ij} - \Delta\bar{E}_{ij}. \end{aligned} \quad (\text{C.29})$$

We also have

$$\begin{aligned} a^2 \delta G_0^0 &= 2[3\mathcal{H}^2 A - 3\mathcal{H}C' + \Delta(C + \mathcal{H}B - \mathcal{H}E') + 3KC], \\ a^2 \delta G_i^0 &= -2D_i [\mathcal{H}A - C' - K(B - E')] - \frac{1}{2}[\Delta + 2K](\bar{E}'_i - \bar{B}_i), \\ a^2 \delta G_j^i &= D^i D_j [(E' - B)' + 2\mathcal{H}(E' - B) - (C + A)] \\ &\quad + \delta_j^i [-\Delta(E' - B)' - 2\mathcal{H}\Delta(E' - B) - 2C'' - 4\mathcal{H}C' + \Delta C + 2KC \\ &\quad + 2\mathcal{H}A' + \Delta A + 2(2\mathcal{H}' + \mathcal{H}^2)A] \\ &\quad + \gamma^{ik} D_{(k} \left\{ [\bar{E}'_{j)} - \bar{B}_{j)}]' + 2\mathcal{H}[\bar{E}'_{j)} - \bar{B}_{j)}] \right\} \\ &\quad + \bar{E}''_{ji} + 2\mathcal{H}\bar{E}'_{ji} - \Delta\bar{E}_j^i + 2K\bar{E}_j^i, \end{aligned} \quad (\text{C.30})$$

whereas the perturbed curvature of the spatial sections is

$$\delta^{(3)}R = -\frac{4}{a^2}(\Delta + 3K)C, \quad (\text{C.31})$$

and the extrinsic curvature tensor for the hypersurfaces of constant conformal time is

$$K_j^i = -\frac{1}{2} \left[\mathcal{H}(1 + \mathcal{K})\delta_j^i + \left(D^i D_j - \frac{1}{3}\delta_j^i \Delta \right) \sigma - 2D^{(i} \bar{\sigma}_{j)} - \bar{\sigma}_j^i \right], \quad (\text{C.32})$$

with

$$\mathcal{K} = -A + \frac{1}{\mathcal{H}} \left(C' + \frac{1}{3}\Delta\sigma \right), \quad \sigma = E' - B, \quad \bar{\sigma}^i = \bar{\Phi}^i, \quad \bar{\sigma}_j^i = \bar{E}_j^i. \quad (\text{C.33})$$

The gauge invariant quantities associated to the geometry are summarized in the table below.

| | | |
|--------|----------------------|---|
| Scalar | Ψ | $= -C - \mathcal{H}(B - E') = -C + \mathcal{H}\sigma$ |
| | Φ | $= A + \mathcal{H}(B - E') + (B - E')' = A - \mathcal{H}\sigma + \sigma'$ |
| | \mathcal{R} | $= C - \mathcal{H} \frac{\delta\varphi}{\varphi'}$ |
| | ζ | $= \Phi + \frac{2}{3} \frac{\Phi' + \mathcal{H}\Phi}{\mathcal{H}(1+w)}$ |
| | ζ_{BST} | $= -C + \frac{\delta\rho}{3(\rho+P)}$ |
| Vector | $\bar{\Phi}_i$ | $= \bar{E}'_i - \bar{B}_i = \bar{\sigma}_i$ |
| Tensor | \bar{E}_{ij} | directly gauge invariant |

C.2.2 Matter

The components of the stress-energy tensor for a perfect fluid, to first order, are

$$\begin{aligned}\delta T_{00} &= \rho a^2 (\delta + 2A), \\ \delta T_{0i} &= -\rho a^2 [(1+w)(D_i v + \bar{v}_i) + D_i B + \bar{B}_i], \\ \delta T_{ij} &= P a^2 \left(h_{ij} + \frac{\delta P}{P} \gamma_{ij} + \pi_{ij} \right),\end{aligned}\tag{C.34}$$

or equivalently

$$\begin{aligned}\delta T_0^0 &= -\rho\delta, \\ \delta T_i^0 &= \rho(1+w)(D_i v + \bar{v}_i + D_i B + \bar{B}_i), \\ \delta T_j^i &= \delta P \delta_j^i + P \pi_j^i,\end{aligned}\tag{C.35}$$

where the anisotropic stress tensor can be SVT-decomposed as

$$\pi_{ij} = \left(D_i D_j - \frac{1}{3} \gamma_{ij} \Delta \right) \pi + D_{(i} \bar{\pi}_{j)} + \bar{\pi}_{ij}.\tag{C.36}$$

Similarly, the scalar pressure is expanded as

$$\delta P = c_s^2 \delta\rho + P\Gamma,\tag{C.37}$$

from which one deduces that

$$\delta w = (c_s^2 - w)\delta + \Gamma w.\tag{C.38}$$

In the special case of a scalar field, we obtain

$$\begin{aligned}\delta T_{00} &= \varphi' \delta \varphi' + 2a^2 V(\varphi) A + a^2 \frac{dV}{d\varphi} \delta \varphi, \\ \delta T_{0i} &= D_i \left[\varphi' \delta \varphi + \left(\frac{\varphi'^2}{2} - a^2 V \right) B \right] + \left(\frac{\varphi'^2}{2} - a^2 V \right) \bar{B}_i, \\ \delta T_{ij} &= \left(\varphi' \delta \varphi' - \varphi'^2 A - a^2 \frac{dV}{d\varphi} \delta \varphi \right) \gamma_{ij} + \left(\frac{\varphi'^2}{2} - a^2 V \right) h_{ij},\end{aligned}\quad (\text{C.39})$$

and therefore

$$\begin{aligned}a^2 \delta T_0^0 &= -\varphi' \delta \varphi' - a^2 \frac{dV}{d\varphi} \delta \varphi + A \varphi'^2, \\ a^2 \delta T_i^0 &= -D_i (\varphi' \delta \varphi), \\ a^2 \delta T_j^i &= \left(\varphi' \delta \varphi - \varphi'^2 A - a^2 \frac{dV}{d\varphi} \delta \varphi \right) \delta_j^i.\end{aligned}\quad (\text{C.40})$$

All the gauge invariant quantities appearing in the description of matter are summarized below.

| | | |
|---------------------|------------------|---|
| Scalar | δ^N | $= \delta + \frac{\rho'}{\rho} (B - E')$ |
| | δ^F | $= \delta - \rho' \frac{C}{H}$ |
| | δ^C | $= \delta + \frac{\rho'}{\rho} (v + B)$ |
| | V | $= v + E'$ |
| | π | |
| | Γ | |
| Vector | \bar{V}_i | $= \bar{v}_i + \bar{B}_i$ |
| | $\bar{\pi}_i$ | |
| Tensor | $\bar{\pi}_{ij}$ | |
| Scalar field | χ | $= \delta \varphi + \varphi' (B - E')$ |
| | Q | $= \delta \varphi - \frac{\varphi' C}{H} = -\varphi' \frac{\mathcal{R}}{H}$ |

Index

- η , 206, 216
- 1+3
 - splitting, 34, 44, 136
- 2dF, 296
- 3-sphere, 722
- Abell
 - 370, 393
 - 2218, 382
- Abundance
 - deuterium, 211
 - helium, 212
 - light elements, 211
- Acceleration of the Universe, 183
- Acoustic oscillation
 - baryon, 298
- Acoustic peak, 321, 367
 - adiabatic perturbations, 322
 - isocurvature perturbations, 322
- Action, 69
 - p -brane, 744
 - classical, 744
 - Dirac, 87
 - Dirac–Born–Infeld, 699
 - Einstein–Hilbert, 39, 738
 - electromagnetism, 42, 78
 - fermion, 108
 - free massive field, 87
 - gauge boson, 107
 - K-essence, 698
 - Nambu–Goto, 745
 - Polyakov, 745
 - Proca, 88
 - scalar field, 43, 71, 457
 - scalar-tensor theory, 580
 - string, 745
- Adiabatic, 272
 - acoustic peaks, 322
 - angular power spectrum, 313, 369
 - initial conditions, 284, 321
 - modes, 281
 - Sachs–Wolfe plateau, 317
- ADM, 44
- Affleck–Dine mechanism, 576
- Age of the Universe, 150, 186, 187
- Algebra
 - Clifford, 88
 - graded, 608
 - Lie, 103, 168
 - Poincaré, 18
- Algorithm
 - Kaiser–Squire, 402
 - Vachaspati–Vilenkin, 666
- $a_{\ell m}$, 312
- Amplification
 - matrix, 381
- Angle
 - Cabbibo, 556
 - deficit, 662
 - deflection, 378
 - weak, 117
 - Weinberg, 117
- Angular
 - distance, 152, 159
 - power spectrum, 311, 354
 - adiabatic, 313
 - Doppler term, 320
 - isocurvature, 313
 - ordinary Sachs–Wolfe, 320
 - reionization, 327
 - scalar modes, 312
 - tensor modes, 315
 - vector modes, 314
- Anisotropic stress
 - radiation, 318
- Anthropic principle, 518, 702
- Anticommutator, 98
- Antiparticle, 98
- Aperture mass, 404
- Approximation
 - flat-sky, 346
 - fluid, 333
 - slow-roll, 459
 - thin-lens, 378
 - tight-coupling, 296, 319
 - WKB, 474
- Arclet, 382
- Asymptotic triviality, 779
- Attraction
 - towards general relativity, 586
- AU, 798
- Axion, 432, 531, 702
- B modes, 407
- Backreaction, 513, 595
- BAO, 298, 365
- Bardeen potential, 254, 269, 284
- Barker theory, 584
- Barotropic fluid, 141
- Baryogenesis, 569

- Electroweak, 573
- Baryon, 100, 292, 364
 - acoustic oscillation, 298, 365
- BBN, 204, 211
 - abundances, 211
 - temperature, 211, 228
- Bessel function, 808
 - spherical, 806, 809
- Bianchi
 - identities, 33, 139
 - space-time, 171
- Bias, 249
 - cosmic, 247
 - magnification, 394
- Big
 - crunch, 148
 - rip, 705
- Binding energy, 207
- Black hole, 419, 426
- Blackbody
 - spectrum, 799
- Bogoliubov coefficients, 600
- Bohr
 - radius, 796
- Bolometric magnitude, 156
- Boltzmann
 - constant, 796
 - equation, 198, 219
 - collisionless, 331
 - perturbed, 330
 - hierarchy, 341, 352
- Bose–Einstein, 189
- Boson
 - exchange, 99
 - gauge, 107
 - Goldstone, 109
 - mass, 110
- Bounce, 148, 785
- BPS limit, 648
- Brackets
 - Lie, 168
 - Poisson, 69
- Brane, 758
 - p , 744
 - Dp , 751
 - gas, 755, 758
- Brans–Dicke
 - parameterization, 580
- Brehmsstrahlung, 328
- Brightness, 334
- Bunch–Davies vacuum, 473, 599, 784
- Cabbibo–Kobayashi–Maskawa matrix, 556
- Calabi–Yau manifold, 755
- CAMB, 356
- Canonical variables, 488
- Cartan classification, 105
- Caustic, 382
- CDM, 365
- Cepheid, 179
- Chaotic inflation, 465
- Chaplygin gas, 700
- Charge
 - central, 608
 - conjugation, 123
- Chemical potential, 192
- Chern–Simons number, 572
- Chirality, 114
- Christoffel symbols
 - definition, 27, 29
 - Friedmann–Lemaître, 134, 817
 - perturbed Friedmann–Lemaître, 819
- Chromodynamics, 113
- C_t , 312, 357
- Classical
 - field, 628
 - mechanics, 69
- Classification
 - Cartan, 105
- Clebsch–Gordan
 - coefficient, 807
- Clifford
 - algebra, 88
 - translation, 718
- Cluster, 419
 - Coma, 420
 - mass, 421
- Clustering
 - stable, 301
- CMB, 216, 221, 307
 - anisotropies, 225
 - Doppler, 311
 - Sachs–Wolfe, 311
 - dipole, 224
 - foregrounds, 225
 - temperature, 217
- CMBEASY, 356
- CMBFAST, 356
- COBE, 1, 217, 221, 225, 231, 358
- CODATA, 803
- Codazzi relation, 37
- Coefficient
 - Bogoliubov, 600
 - Clebsch–Gordan, 807
- Coincidence problem, 233
- Coleman–Mandula theorem, 606
- Coleman–Weinberg
 - correction, 680
- Collapse
 - spherical, 300
- Colour, 113
- Commutator, 80
- Comoving
 - observer, 136
 - radial distance, 151
- Compactification, 739, 743, 754
 - warped, 764
- Compton
 - distortion, 223
 - inverse scattering, 328

- Conditions
 - Dirichlet, 750
 - junction, 485
 - mixed, 751
 - Neumann, 749
 - Sakharov, 569
- Conformal
 - coupling, 596
 - invariance, 596
 - time, 140, 458
- Connectedness, 649
- Conservation
 - $B - L$, 573
 - equation, 41, 139, 199
 - law, 73
- Consistency relation, 496, 499, 500, 504, 528
- Constant
 - Boltzmann, 796
 - cosmological, 275, 366, 684
 - coupling, 107
 - Fermi, 796
 - fine structure, 713, 796
 - fundamental, 58
 - gravitation, 582, 715
 - Hubble, 150, 365, 799
 - Newton, 796
 - Planck, 796
 - structure, 103
 - varying, 711
- Continuity equation, 241
- Convergence, 380, 397, 399
 - power spectrum, 399
- Coordinates
 - generalized, 69
- Copernican principle, 130, 704
- Correlation function, 246, 812
 - angular, 311
- Correlation length, 635, 637, 638
- Cosmic
 - bias, 247
 - distribution function, 247
 - error, 248
 - variance, 248, 357
- Cosmic microwave background, 216, 221
 - anisotropies, 225
 - dipole, 224
 - foregrounds, 225
- Cosmic shear, 380, 397, 402
 - measurement, 405
- Cosmic strings, 641, 661
 - global, 646
 - kink, 667
 - Nambu-Goto, 646
 - non-Abelian, 671
 - simulation, 666
 - superconducting, 646
 - equation of state, 646
 - state parameter, 646
- Cosmological
 - constant, 184, 228, 366, 684
- problem, 233, 685
- model, 2, 129
- parameters, 364
- principle, 2, 130, 167, 229
- space-time, 131
- Coupling
 - conformal, 596
 - minimal, 596
- Covariant derivative, 28
- Covering space, 718
- CP violation, 556, 570, 573
- CPT theorem, 124
- Critical
 - exponent, 637
 - lens, 382
 - line, 382, 386
 - temperature, 630
- Cross-section, 388
- Curvaton, 528, 532, 781
- Curvature, 366
 - extrinsic, 37, 45, 486, 820
 - perturbation, 484
 - scalar, 33
- Cylinder condition, 739
- D term, 614
- Dp -brane, 751
- d'Alembertian, 30
- D-particle, 751
- Dark energy, 185, 233, 592, 684, 687
 - equation of state, 185
- Dark matter, 201, 203, 233, 292, 365, 393, 411
 - candidates, 430, 438
 - cold, 424, 431
 - halo model, 386
 - hot, 427, 431
 - repulsive, 428
 - solid, 697
 - strongly self-interacting, 428
 - warm, 427, 431
- DBI, 699
- de Sitter, 455
 - massive field, 476
 - test field, 470
- de Vaucouleur law, 180
- Deceleration parameter, 158, 183
- Decomposition
 - SVT, 52, 251, 257, 738
- Decoupling, 190, 194, 217, 220
- Deficit angle, 662
- Deflection angle, 378
- Deformation matrix, 397
- Degrees of freedom
 - relativistic, 191, 196
- Derivative
 - covariant, 28, 30
 - functional, 70
 - Lie, 31
 - spatial, 35

- Deuterium, 211, 215
 - mass, 798
- Dilaton, 592, 741, 778
- Dirac
 - equation, 87
 - spinor, 606
- Distance
 - angular, 152, 159
 - comoving, 152, 159
 - comoving radial, 151, 158
 - duality, 155
 - luminosity, 154, 159, 182
 - modulus, 155
- Distortion
 - matrix, 380
 - spectral, 328
- Distribution function, 189, 330
- Divergences
 - infra-red, 475
- Duality, 752
 - scale factors, 778
 - T, 755
- E modes, 407
- Effect
 - Kaiser–Stebbins, 665
 - Sunyaev–Zel'dovich, 181, 328
- e-fold, 452, 461, 464, 502, 532, 539
 - chaotic inflation, 465
- Eigenmodes
 - Laplacian, 813
- Einstein
 - equations, 38, 40
 - brane-induced, 761
 - gauge invariant, 260
 - perturbations, 260
 - semi-classical, 595
 - equivalence principle, 20
 - frame, 581, 742
 - linearized equation, 49
 - radius, 376, 382, 388
 - tensor, 33, 817
 - Friedmann–Lemaître, 135
- Einstein–Hilbert
 - action, 39, 45
- Einstein–Maxwell theory, 741
- Ekpyrotic, 781
- Electric field, 43
- Electrodynamics
 - scalar, 108
- Electromagnetic tensor, 42
- Electromagnetism, 42, 78
 - spectrum, 799
- Electron mass, 798
- Electroweak force, 114
- Embedding, 455
- Energy condition, 36
 - strong, 162
 - weak, 163
- Entropy, 193, 222, 280, 284, 821
- Equality, 188, 282
 - temperature, 228
- Equation
 - Boltzmann, 219
 - collisionless, 331
 - gauge invariant, 341
 - perturbed, 330
 - polarization, 344
 - collision term, 338
 - Codazzi, 761
 - conservation, 41, 139, 238
 - scalar modes, 265
 - continuity, 238, 241
 - Dirac, 87
 - Einstein, 38, 40
 - linearized, 49, 52
 - scalar modes, 263
 - semi-classical, 595
 - tensor modes, 262
 - vector modes, 263
 - Euler, 41, 238
 - scalar modes, 265
 - vector modes, 265
 - Euler–Lagrange, 69
 - flow, 506
 - Fokker–Planck, 520
 - Friedmann, 139, 818
 - brane, 767
 - dynamical system, 145
 - generalized, 42, 140
 - reduced, 141
 - scalar field, 458
 - gauge invariant, 260
 - geodesic deviation, 47
 - geodesics, 26
 - linearized, 51
 - gravitational lensing, 378
 - Hamilton–Jacobi, 458
 - Helmholtz, 727
 - Jacobi, 461
 - Killing, 33
 - Klein–Gordon, 44, 86, 458
 - inflation, 481
 - slow-roll, 478
 - Kolmogorov, 522
 - Langevin, 519
 - lens, 375
 - Liouville, 331
 - Mathieu, 511
 - Maxwell, 43, 88
 - Mészáros, 290
 - of state, 140, 271, 818
 - dark energy, 185
 - parameterization, 705
 - perturbed geodesic, 308
 - Poisson, 239
 - test, 65
 - Raychaudhuri, 36, 139, 172
 - Sachs, 395
 - Saha, 217

- Schrödinger, 83
- Tomonaga–Schwinger, 551
- Weyl, 115
- Equivalence principle of, 14
- EROS, 390
- Estimator, 248
- Euclidean space, 718
- Euler
 - B function, 804
 - Γ function, 804
 - equation, 41, 238
- Euler–Lagrange equation, 69
- Eulerian perturbations, 241
- Event horizon, 160
- Extrinsic curvature, 37, 45, 486, 820
- f_{NL} , 530
- F term, 612
- Faber–Jackson relation, 418
- Fayet–Iliopoulos, 617
- Fermi constant, 796
- Fermi–Dirac, 189
- Fermion mass, 110
- Field canonically conjugate, 72
 - electric, 43
 - Kalb–Ramond, 751
 - magnetic, 43
 - scalar, 43, 87
 - complex, 95
- Fifth force, 58
- Filter, 248
- Filtering, 247
- Fixed point, 169
- Flat-sky, 346
- Flatness, 451, 452
 - problem, 229
- Flow irrotational, 38
- Fluid approximation, 333
 - barotropic, 141
 - perfect, 41
- Fock representation, 490
 - space, 93
- Fokker–Planck equation, 520
- Force fifth, 58
- Foreground emissions, 225
- Formula Sachs–Wolfe, 307
- Fourier transform, 806, 812
 - spherical, 814
- Frame Einstein, 581, 742
- Jordan, 580, 741
 - string, 580, 741
- Freeze-out, 190, 194, 200
 - weak interaction, 208
- Friedmann equation, 139, 818
 - dynamical system, 145
 - generalized, 42, 140
 - reduced, 141
 - scalar-tensor, 584
- Friedmann–Lemaître metric, 133
- Function Bessel, 808
 - correlation, 246
 - distribution, 330, 339
 - gauge invariant, 339
 - Euler B, 804
 - Euler Γ , 804
 - Green, 320, 521
 - Hankel, 808
 - spherical Bessel, 806, 809
 - transfer, 292
- Functional derivative, 70
- Fundamental form, 485
 - observer, 136
 - plane, 179, 418
- Galaxy catalogue, 296
 - elliptical, 417
 - Milky Way, 414
 - rotation curves, 65
 - spiral, 414, 415
- Galilean group, 13
- Gauge boson, 107
 - comoving, 260
 - dependence, 273
 - flat-slicing, 259
 - freedom, 50
 - Gauss, 259
 - group, 107
 - harmonic, 50
 - invariant distribution function, 339
 - invariant perturbations, 251
 - invariant variables, 53, 253, 254, 339, 480, 820
 - longitudinal, 258
 - Lorentz, 89
 - Newtonian, 258
 - normal Gaussian, 765
 - problem, 252
 - symmetry, 107
 - synchronous, 259
 - variables, 257
 - Gauss relation, 37, 762
 - scalar relation, 37

- Gauss–Bonnet
 - term, 769
- Gaussian
 - filter, 248
 - initial conditions, 473
- Gegenbauer
 - polynomials, 805
- Gell–Mann–Nishijima
 - relation, 116, 120, 562
- General relativity, 11
 - D dimensions, 737
 - modification, 443
- Generalized
 - coordinates, 69
 - momenta, 69
 - velocity, 69
- Generator, 103
- Geodesic
 - deviation equation, 47
 - equation, 26
 - linearized, 51
 - perturbed, 308
 - flow, 35
- GeV, 797
- Goldstone boson, 109
- Grand Unification
 - E_6 , 563
 - $SO(10)$, 563
 - $SU(5)$, 563
- Grassmann variables, 607
- Gravitational lensing, 378
 - by strings, 663
 - thin-lens, 375
- Gravitational waves, 53, 54
 - CMB, 368
 - detection, 64
 - energy density, 494
 - inflation, 482
 - polarization, 54
 - power spectrum, 329, 493
 - primordial, 368
 - propagation, 53, 262, 267
 - quantization, 492, 496, 498
 - radiation, 63
- Gravitino, 434, 509
 - definition, 622
 - problem, 434
 - Rarita–Schwinger action, 622
- Gravity
 - induced, 771
 - quantum, 748
- Green function, 320, 521
- Group, 101
 - Abelian, 102
 - dimension, 106
 - exceptional, 105
 - fundamental, 649
 - Galilean, 13
 - gauge, 107
 - Grand Unification, 560
- holonomy, 718
- homotopy, 650
- Lie, 102, 103, 168
- order, 105, 106
- Poincaré, 17
- quotient, 633, 649
- rank, 105, 106
- representation, 104
- simple, 105
- structure constant, 103
- Growth factor, 243
- Hadron, 100
- Halo
 - dark matter, 386
- Hamilton–Jacobi equation, 458
- Hamiltonian, 69, 72
 - density, 45
 - formalism, 44
- Hankel functions, 808
- Harmonic
 - gauge, 50
 - oscillator, 83
- Harmonics
 - spherical, 335
 - of spin s , 344
- HDM, 427
- Heisenberg
 - representation, 82, 489
- Helium, 212, 215
- Helmoltz equation, 727
- Hierarchical model, 299
- Hierarchy problem, 605
- Higgs
 - mechanism, 108
 - potential, 108
- Hilbert space, 79
- Homeomorphism, 649
- Homogeneity, 131, 170
- Horizon
 - event, 160
 - particle, 162
 - problem, 229, 453
 - sound, 637
- Hubble
 - constant, 150, 178, 181, 188, 228, 365, 392, 799
 - diagram, 177, 178, 185
 - distance, 150
 - flow, 138
 - law, 137, 177
 - generalized, 47
 - time, 150
- Hybrid inflation, 468
- Hypercharge, 116
- Hypothesis
 - cosmological, 129
 - large numbers, 711
- HZT, 183

- Identities
 - Bianchi, 33
- Identity
 - Gauss, 762
 - Weingarten, 761
- Images
 - multiple, 376
- Index
 - spectral, 368, 478, 496, 499
- Induced metric, 485
- Inflation, 451
 - D -term, 683
 - F -term, 681
 - assisted, 524
 - attractor, 463
 - chaotic, 465
 - consistency relation, 496, 499, 500, 504, 528
 - constraints, 503
 - end, 463, 484, 487, 507
 - eternal, 516
 - flow equations, 506
 - gauge invariant variables, 480
 - gravity waves, 482
 - hybrid, 468, 661, 684
 - isocurvature modes, 524, 526
 - models, 678
 - modulated fluctuations, 529
 - multifield, 523
 - new, 455
 - Nflation, 525
 - old, 454
 - phase space, 463
 - power-law, 467, 495
 - single field, 457
 - slow-roll, 680
 - stochastic, 518
 - super, 776
 - supergravity, 678
- Inflaton, 454
 - decay rate, 509
 - potential, 505
 - quantum fluctuations, 478
- Initial condition
 - chaotic, 466
- Instanton, 572
- Intercommutation, 666
- Internal symmetry, 97
- Invariance
 - conformal, 596
 - discrete, 121
- Invariant variety, 169
- Island Universe, 517
- Isocurvature, 524, 526
 - acoustic peaks, 322
 - angular power spectrum, 313, 369
 - correlated modes, 527
 - initial conditions, 285, 322
 - modes, 281
 - Sachs–Wolfe plateau, 318
- Isospin, 116
- Isothermal
 - sphere, 385
- Isotropy, 131, 171
 - group, 169
- Jacobi
 - equation, 461
 - identity, 168
- Jeans length, 239
- Jordan frame, 580, 741
- Junction conditions, 485, 762
 - cosmology, 486
- K-correction, 157
- K-essence, 698
- Kähler potential, 622
- Kaiser–Squire algorithm, 402
- Kaiser–Stebbins effect, 665
- Kalb–Ramond field, 751
- Kaluza–Klein, 737
- κ , 51
- Killing
 - equation, 33
 - vector, 33, 168
 - Friedmann–Lemaître, 135
- Klein–Gordon
 - equation, 44, 86
- Kolmogorov equation, 522
- Lagrangian, 69, 71
 - density, 45
 - perturbations, 241
- Landscape, 518
- Langevin equation, 519
- Laplacian
 - eigenmodes, 813
- Lapse function, 44
- Large-scale structure, 296
 - numerical simulation, 302
- Last-scattering surface, 220
 - thickness, 324
- Law
 - de Vaucouleur, 180
 - Hubble, 137, 177
- Legendre
 - polynomials, 805, 811
- Leibniz rule, 70
- Lemma
 - Stewart–Walker, 255
- Length
 - correlation, 635, 637, 638
 - Jeans, 239
 - Planck, 797
- Lens
 - critical, 382
 - equation, 378
 - space, 725
 - strong, 382
- Lensing

- strong, 382
- weak, 382
- Lepton, 100, 115
- Lie
 - algebra, 103, 168
 - derivative, 31
 - group, 102, 168
- Lifetime
 - neutron, 798
 - proton, 567
- Light
 - bending, 60
 - curve, 388
- Limber approximation, 399
- Limit
 - BPS, 648
- Liouville
 - equation, 331
 - operator, 198
- Lithium, 215
- Look-back time, 151, 158
- Lorentz
 - gauge, 89
 - invariance, 20, 58
 - tests, 58
 - invariant measure, 89
 - transformations, 14
- Low surface brightness, 418
- LSP, 433, 606, 621
- Luminosity, 156
 - distance, 154, 159, 182
 - solar, 796
- M-theory, 748
- MACHO, 390
- Magnetic field, 43
- Magnification, 376
 - bias, 394
- Magnitude, 156
 - bolometric, 156
- Majorana
 - spinor, 606
- Manifold, 21
 - Calabi–Yau, 755
- Mass
 - aperture, 404
 - cluster, 421
 - dynamical, 413
 - electron, 798
 - fraction, 207
 - lens, 414
 - neutrinos, 559, 568
 - Planck, 797
 - proton, 798
 - sheet, 386
 - solar, 796
- Matching conditions, 762
- Mathieu equation, 511
- Matrix
 - amplification, 381
- Cabbibo–Kobayashi–Maskawa, 556
- CKM, 555
- deformation, 397
- distortion, 380
- optical, 396
- Pauli, 88
- symplectic, 106
- Matter–radiation equality, 188
- Maxwell
 - equation, 43, 88
 - theory, 42
- Measure
 - Lorentz invariant, 89
- Mechanics
 - classical, 69
 - quantum, 79
- Mechanism
 - 'see-saw', 575
 - Brandenberger–Vafa, 757
 - Higgs, 108
 - Kibble
 - cosmic strings, 641
 - domain walls, 638
 - Mercury perihelion, 62
 - Meson, 100
 - Mészáros equation, 290
 - Metric
 - 5-dimensional, 738
 - conical, 663
 - induced, 485, 761
 - Kaluza–Klein, 738
 - MeV, 797
 - Microlensing, 387
 - Milky Way, 414
 - Milne Space-time, 144
 - Model
 - Big-Bang, 177
 - hierarchical, 299
 - particle physics, 68
 - Pati–Salam, 562
 - Randall–Sundrum, 764
 - spherical collapse, 300
 - standard cosmological, 129
 - Swiss-cheese, 171
 - Universe, 129
 - Modulated fluctuations, 529
 - Modulus, 755
 - Momenta
 - generalized, 69
 - MOND, 440
 - Monopoles, 231, 647, 656
 - problem, 232
 - MSSM, 618
 - Mukhanov–Sasaki
 - variables, 480, 488
 - Multigravity, 701
 - Multipolar expansion, 335, 348
 - N-body, 302
 - Nambu–Goto

- action, 745
- Neutralino, 434
- Neutrino, 196, 295, 431
 - Dirac, 568
 - Majorana, 568
 - mass, 559, 568
 - oscillations, 557
 - sterile, 432
- Neutron
 - lifetime, 798
- Newton
 - constant, 796
- Newtonian perturbations, 238
- NFW, 425
- Nielsen–Olesen
 - configuration, 644
- Noether theorem, 73
- Non-Gaussianity, 530
- Nuclear equilibrium, 206
- Nucleation, 636
- Nucleosynthesis, 204, 211
- Number counts, 157
- Numerical simulation, 302
- Observer
 - fundamental, 136
- Oklo, 713
- On shell, 90
- Operator
 - annihilation, 84
 - creation, 84
- Optical
 - depth, 220
 - matrix, 396
- Optics
 - geometric, 48
- Parameter
 - cosmological, 141, 364
 - deceleration, 158
 - post-Newtonian, 584
 - constraints, 62
 - definition, 60
 - slow-roll, 459, 469
 - Stokes, 343
- Parity, 381
- Parsec, 798
- Particle
 - creation, 491
 - horizon, 162
- Pati–Salam
 - model, 562
- Pauli matrices, 88
- Pauli–Lubanski spin vector, 609
- p*-brane, 744
- PDF, 248
- Penrose diagrams, 163
- Perfect fluid, 255
 - stress-energy tensor, 138, 817, 821
- Perturbations
 - Eulerian, 241
 - gauge invariant, 251
 - growth factor, 243
 - Lagrangian, 241
 - Newtonian, 238
 - non-linear regime, 249, 299
 - Phase transition, 454, 628
 - first-order, 635
 - order, 632
 - second-order, 636
 - Photon, 117
 - fluid, 294
 - Photon–axion mixing, 702
 - Photon–baryon plasma, 318
 - Planck
 - constant, 796
 - length, 797
 - mass, 797
 - temperature, 797
 - time, 797
 - units, 797
 - Plasma
 - photon–baryon, 318
 - Pocket Universe, 517
 - Poincaré
 - algebra, 18
 - group, 17
 - Poisson
 - brackets, 69, 73
 - equation, 65, 239
 - Polarization, 342, 407
 - Boltzmann equation, 344
 - Polyhedron
 - fundamental, 718
 - Polynomials
 - Gegenbauer, 805
 - Legendre, 805, 811
 - Pontryagin index, 643
 - Position invariance, 20
 - tests, 58
 - Post-Newtonian, 60, 584
 - Potential
 - effective, 631, 632, 682
 - Higgs, 108
 - Kähler, 622
 - projected, 380
 - Power spectrum, 247, 282, 474, 492
 - angular, 311, 354
 - scalar modes, 312
 - blue, 478
 - convergence, 399
 - gravity waves, 493
 - non-linear regime, 301
 - normalization, 812
 - observation, 298
 - red, 478
 - scale invariant, 478
 - Power-law inflation, 467
 - Pre-Big Bang, 775
 - Preheating, 507, 510

- Primordial nucleosynthesis, 204
 Principle
 anthropic, 518, 702
 Copernican, 130, 704
 cosmological, 130, 167
 equivalence, 19
 Einstein, 20
 tests, 55
 weak, 14, 20
 superposition, 80, 550
 uncertainty, 79
 uniformity, 130
 unitarity, 82
 Probability distribution function, 248
 Problem
 η , 683
 coincidence, 233
 cosmological constant, 233, 685
 dark sector, 233
 flatness, 229, 451, 452
 gauge, 252
 gravitino, 434
 hierarchy, 605, 759
 horizon, 229, 453
 moduli, 755
 monopoles, 232, 656
 of the origin of structures, 231
 Polonyi, 755
 standard Big-Bang, 452
 trans-Planckian, 532
 Proca theory, 88
 Profile
 β , 328
 NFW, 425
 Sersic, 418
 universal, 416
 Proton
 decay, 567
 lifetime, 206
 mass, 798
 QCD, 113
 Lagrangian, 113
 Quantization
 density perturbation, 488
 gravity waves, 492, 496, 498
 slow-roll, 494, 496
 Quantum field
 de Sitter, 470, 476
 inflaton, 478
 slow-roll, 477
 Quantum gravity, 570
 Quarks, 114
 Quasar, 390
 absorption spectra, 714
 Quintessence, 592, 690
 extended, 695
 problems, 696
 \mathcal{R} , 271, 484
 R -parity, 620
 Radiation
 anisotropic stress, 318
 Radion, 764
 Raychaudhuri
 equation, 36, 139, 172
 RDM, 428
 Recession velocity, 137
 Reciprocity theorem, 153
 Recombination, 217, 218
 Reconnection, 666
 Redshift, 48, 137, 308
 time drift, 159
 Regge slope, 745
 Reheating, 507
 parametric, 510
 tachyonic, 515
 temperature, 509
 Reionization, 327
 Relation
 Codazzi, 37
 Faber–Jackson, 418
 Gauss, 37, 762
 Gauss, scalar, 37
 Gell-Mann–Nishijima, 116, 562
 Tully–Fisher, 179, 416
 Relativity
 general, 19
 tests, 55, 60, 65
 special, 14
 Relics, 231
 cold, 201
 hot, 203
 thermal, 431
 warm, 204
 Representation
 adjoint, 104
 chiral, 88
 conjugate, 104
 equivalent, 104
 fundamental, 106
 Heisenberg, 82
 irreducible, 104
 Schrödinger, 82
 unitary, 104
 Resonance
 parametric, 510
 stochastic, 513
 Ricci
 tensor, 33, 817
 Friedmann–Lemaître, 135
 Riemann
 tensor, 31, 817
 Rotation curves, 65, 412
 Runaway dilaton, 592
 Running, 500
 Rydberg constant, 796
 Sachs equation, 395
 Sachs–Wolfe

- effect, 307
- formula, 311
- integrated, 326
- plateau, 317, 501
- Saha equation, 217
- Sakharov conditions, 569
- Scalar
 - modes, 263, 265
 - propagation, 269
- Scalar field, 43, 87
 - action, 457
 - baryonic, 576
 - D -dimensional, 596
 - Friedmann equation, 458
 - inflaton, 457
 - quantization, 495
 - stress-energy tensor, 479, 818
- Scalar-tensor
 - theory, 579
- Scalar-tensor theory
 - Einstein equation, 581
 - Klein–Gordon equation, 581
- Scale factor, 140
 - slow-roll, 477
- Scaling regime, 667
- SCDM, 424
- Scenario
 - ekpyrotic, 781
- Schrödinger
 - equation, 83
 - representation, 82
- Schwarzschild
 - radius, 796
 - space-time, 60
- SCP, 183
- SDSS, 296
- See-saw mechanism, 568
- Self-reproduction, 516
- Semi-topological defects, 651
- Sersic profile, 418
- Shapiro effect, 61
- Shear
 - cosmic, 380, 397, 402
 - measurement, 405
- Shell crossing, 242
- Shift vector, 44
- Sidereal year, 798
- SIDM, 428
- σ_8 , 249
- Silk damping, 323
- Skewness, 250
- Slow-roll, 494
 - approximation, 459
 - attractor, 463
 - expansion of the potential, 461
 - Klein–Gordon, 478
 - parameters, 459, 469
 - predictions, 496
 - quantization, 498
 - scale factor, 477
- test field, 477
- Smoothing, 247
- sneutrino, 433
- Solar
 - luminosity, 796
 - mass, 796
- $SO(n)$, 106
- $SO(10)$, 564
- Sound
 - speed, 141, 239, 818, 821
- Space
 - chimney, 721
 - connected, 649
 - constant curvature, 34
 - covering, 718
 - cylindrical, 739
 - Euclidean, 718
 - Fock, 93
 - Hantzsch–Wendt, 720
 - Hilbert, 79
 - hyperbolic, 726
 - Klein, 721
 - lens, 725
 - quotient, 634
 - slab, 722
 - spherical, 722
- Space-time
 - Minkowski, 16
 - Bianchi, 171
 - classification, 172
 - type I, 172
 - conical, 663
 - cosmology, 131
 - curved, 19, 594
 - de Sitter, 166, 455
 - Friedmann–Lemaître, 133, 816
 - Christoffel symbol, 134
 - dynamics, 143
 - Einstein tensor, 135
 - Killing vector, 135
 - perturbed, 251, 819
 - Ricci tensor, 135
 - Lemaître–Tolman–Bondi, 170
 - maximally symmetric, 34, 132
 - Milne, 144
 - Minkowski, 15, 164
 - Newtonian, 11
 - Schwarzschild, 60
 - Swiss-cheese, 171
- Special relativity, 14
- Spectral
 - distortion, 328
 - index, 368, 478, 496
 - running, 500
 - slow-roll, 499
- Spectrum
 - blackbody, 799
 - electromagnetic, 799
- Speed
 - of light, 796

- of sound, 239
- Sphaleron, 572
- Sphere
 - isothermal, 385
- Spherical
 - Bessel function, 809
 - Bessel functions, 806
 - collapse, 300
 - harmonics, 805
 - waves, 93
- Spinodal instability, 468
- Spinor
 - Dirac, 606
 - Majorana, 606
 - Weyl, 606
- Spiral galaxy, 414, 415
- Splitting
 - $1+3$, 34
- $Sp(2n)$, 106
- Stable clustering, 301
- Statefinder, 707
- Sterile neutrino, 432
- Stewart–Walker lemma, 255
- Stochastic description, 246
- Stokes parameters, 343
- Stress-energy
 - tensor, 77
 - definition, 40
 - electromagnetism, 42
 - perfect fluid, 255
 - scalar field, 43
 - symmetrization, 78
- String
 - bosonic, 753
 - closed, 749
 - cosmic, 661
 - cosmology, 737
 - frame, 580, 741
 - heterotic, 751
 - open, 749
 - theory, 737
 - type I, 749
 - type IIA and IIB, 751
- Strong
 - energy condition, 452
 - lensing, 382
- Sub-Hubble, 266, 289
- SUGRA, 624
- $SU(n)$, 106
- Sunyaev–Zel'dovich effect, 181, 328
- Super-Hubble, 266, 285
- Superfields, 612
- Supergravity, 678
- Supermultiplet, 609
 - vector, 614
- Supernovæ Ia, 179
- Superposition principle, 80
- Superpotential, 612
- Superstring theory, 747
- Supersymmetry, 678
- extended, 609
- multiplet, 609
- Surface of transitivity, 169
- SVT, 52, 251
 - decomposition, 257, 335, 481, 738, 819
- Symmetry
 - breaking, 111, 118, 628, 630
 - breaking scheme, 633
 - gauge, 107
 - internal, 97
 - local, 107
 - spontaneous breaking, 632
- Symmetry breaking, 628, 630
 - Abelian–Higgs model, 630
 - grand unification, 564
- 't Hooft–Polyakov monopole, 647
- T-duality, 755
- Tachyon, 699, 746
- Temperature, 334
 - BBN, 211, 228
 - CMB, 217, 228, 799
 - critical, 630
 - decoupled species, 195
 - decoupling, 799
 - equality, 228
 - freeze-out, 206
 - Hagedorn, 632, 756
 - neutrinos, 799
 - of the Universe, 190, 195
 - Planck, 797
 - reheating, 509
- Tensor
 - calculus, 25
 - definition, 23
 - Einstein, 33
 - Friedmann–Lemaître, 817
 - perturbed Friedmann–Lemaître, 820
 - electromagnetic, 42
 - extrinsic curvature
 - perturbed Friedmann–Lemaître, 820
 - modes, 261, 262, 267
 - Ricci, 33
 - Friedmann–Lemaître, 817
 - perturbed Friedmann–Lemaître, 819
 - Riemann, 31
 - Friedmann–Lemaître, 817
 - stress-energy, 77, 255
 - canonical, 77
 - complex field, 96
 - definition, 40
 - electromagnetism, 42
 - perfect fluid, 138, 817, 821
 - scalar field, 43, 821
- Term
 - Gauss–Bonnet, 769
- Test
 - equivalence principle, 55
 - general relativity, 55, 60, 65
 - Lorentz invariance, 58

- position invariance, 58, 712
- universality of free fall, 14, 55, 712
- Textures, 648
- Theorem
 - Coleman–Mandula, 606
 - CPT, 124
 - Haag–Lopuszanski–Sohnius, 606
 - Noether, 73
 - reciprocity, 153
 - virial, 419
- Theory
 - Barker, 584
 - Einstein, 39
 - Maxwell, 42, 88
 - Proca, 88
 - scalar-tensor, 579
 - supergravity, 678
 - superstring, 747
- Thermal relics, 431
- Thermodynamics, 188
 - at equilibrium, 188
 - out-of-equilibrium, 197
 - superstrings, 756
- Thomson
 - scattering, 338, 348
 - polarization, 342
- Tight coupling approximation, 296, 319
- Time
 - conformal, 140
 - delay, 383
 - Hubble, 799
 - look-back, 151, 158
 - Planck, 797
 - quench, 637
- Tomonaga–Schwinger
 - equation, 551
- Top-hat, 248
- Topological defects, 531, 634, 697
 - cosmic strings, 661
 - domain walls, 653
 - hybrid, 660
 - scaling regime, 667
 - texture, 648
- Topology
 - of the Universe, 716
- Torus, 719
- Trans-Planckian, 532
- Transfer function, 292
- Transformation
 - conformal, 581, 596
- Transition
 - electroweak, 574
 - phase, 628
- Translation
 - Clifford, 718
- Transport
 - parallel, 29
- Tully–Fisher relation, 179, 416
- $U(n)$, 106
- Uncertainty principle, 79
- Uniformity principle, 130
- Unitarity
 - bound, 435
 - principle, 82
- Unitary representation, 104
- Units
 - conventions, 142
 - natural, 797
 - Planck, 797
- Universality of free fall, 14, 20, 55
 - tests, 14, 55, 712
- Universe model, 129
- Vachaspati–Vilenkin
 - algorithm, 666
- Vacuum, 597
 - Bunch–Davies, 473, 599
 - energy, 85, 685
 - state, 490
 - stress-energy tensor, 685
- Variables
 - canonical, 488
 - gauge invariant, 53, 253
 - Grassmann, 607
 - Mukhanov–Sasaki, 488
- Vector, 22
 - Killing, 33
 - modes, 261, 263, 265, 267
 - shift, 44
- Velocity
 - generalized, 69
 - proper, 137
- VEV, 108, 629
- Virial theorem, 419
- Vortex, 640
- Vorton, 438, 646
- Waves
 - gravitational, 53
 - primordial, 482
 - spherical, 93
- WDM, 427
- Weak
 - angle, 117
 - interaction, 208
 - lensing, 382
- Weinberg angle, 117
- Weyl
 - equation, 115
 - spinor, 115, 606
- Wigner symbols, 807
- WIMP, 424, 438
- Wimpzilla, 435
- Winding number, 643
- Window function, 248
- WKB approximation, 474, 493, 535
- WMAP, 1, 184, 187, 216, 221, 358, 502, 504, 531, 732
- W^\pm , 117

Wronskian, 490, 599

potential, 583

Y_p , 205, 212, 215

Yukawa

coefficients, 120

Z^0 , 117

ζ , 270, 484

ζ_{BST} , 484, 487





5410353283

This book provides an introduction to cosmology and its observational methods. It will help students to get into research by providing definitions and main techniques and ideas discussed today. The book is divided into three parts. Part 1 summarises the fundamentals in theoretical physics needed in cosmology (general relativity, field theory, particle physics). Part 2 describes the standard model of cosmology and includes cosmological solutions of Einstein equations, the hot big bang model, cosmological perturbation theory, cosmic microwave background anisotropies, lensing and evidence for dark matter, and inflation. Part 3 describes extensions of this model and opens up current research in the field: scalar-tensor theories, supersymmetry, the cosmological constant problem and acceleration of the universe, topology of the universe, grand unification and baryogenesis, topological defects and phase transitions, string inspired cosmology including branes and the latest developments. The book provides details of all derivations and leads the student up to the level of research articles.

Patrick Peter and **Jean-Philippe Uzan** are at the Institut d'Astrophysique de Paris, France.

'Fills a niche that other recent cosmology texts leave open, namely self-contained derivations in cosmology that span both fundamental issues and applications to the real universe that are of great interest to observers.'

Joseph Silk, *University of Oxford*

'A remarkable book. Written with great authority and enthusiasm, it gives a comprehensive view of primordial cosmology today. The style is accessible to a novice for a good introduction, as well as being technically precise enough to be useful for specialists.'

Ted Jacobson, *University of Maryland*

ALSO AVAILABLE FROM
OXFORD UNIVERSITY PRESS

Cosmology

Steven Weinberg

Particle Astrophysics: Second Edition

Donald H. Perkins

Cosmic Anger—Abdus Salam: The First Muslim Nobel Scientist

Gordon Fraser

Revolutionaries of the Cosmos—The Astro-Physicists

Ian S. Glass

Cover image: 'Le vent' by Madeleine Attal (1997).

OXFORD
UNIVERSITY PRESS

www.oup.com

ISBN 978-0-19-920991-0



9 780199 209910