Georg Wolschin
Editor

LECTURE NOTES IN PHYSICS 800

Lectures on Cosmology

Accelerated Expansion of the Universe

Springer

Lecture Notes in Physics

The Lecture Notes in Physics

The series Lecture Notes in Physics (LNP), founded in 1969, reports new developments in physics research and teaching – quickly and informally, but with a high quality and the explicit aim to summarize and communicate current knowledge in an accessible way. Books published in this series are conceived as bridging material between advanced graduate textbooks and the forefront of research and to serve three purposes:

- to be a compact and modern up-to-date source of reference on a well-defined topic
- to serve as an accessible introduction to the field to postgraduate students and nonspecialist researchers from related areas
- to be a source of advanced teaching material for specialized seminars, courses and schools

Both monographs and multi-author volumes will be considered for publication. Edited volumes should, however, consist of a very limited number of contributions only. Proceedings will not be considered for LNP.

Volumes published in LNP are disseminated both in print and in electronic formats, the electronic archive being available at springerlink.com. The series content is indexed, abstracted and referenced by many abstracting and information services, bibliographic networks, subscription agencies, library networks, and consortia.

Proposals should be sent to a member of the Editorial Board, or directly to the managing editor at Springer:

Christian Caron
Springer Heidelberg
Physics Editorial Department I
Tiergartenstrasse 17
69121 Heidelberg / Germany
christian.caron@springer.com

G. Wolschin (Ed.)

Lectures on Cosmology

Accelerated Expansion of the Universe

Springer

Georg Wolschin
Universität Heidelberg
Inst. Theoretische Physik
Philosophenweg 19
62190 Heidelberg
Germany
g.wolschin@thphys.uni-heidelberg.de

Wolschin, G. (Ed.), *Lectures on Cosmology: Accelerated Expansion of the Universe*, Lect. Notes Phys. 800 (Springer, Berlin Heidelberg 2010),
DOI 10.1007/978-3-642-10598-2

Lecture Notes in Physics ISSN 0075-8450 e-ISSN 1616-6361
ISBN 978-3-642-10597-5 e-ISBN 978-3-642-10598-2
DOI 10.1007/978-3-642-10598-2
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010920542

Cover design: Integra Software Services Pvt. Ltd., Pondicherry

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The lectures that four authors present in this volume investigate core topics related to the accelerated expansion of the Universe. Accelerated expansion occured in the very early Universe – an exponential expansion in the inflationary period 10^{-36} s after the Big Bang. This well-established theoretical concept had first been proposed in 1980 by Alan Guth to account for the homogeneity and isotropy of the observable universe, and simultaneously by Alexei Starobinski, and has since then been developed by many authors in great theoretical detail.

An accelerated expansion of the late Universe at redshifts $z < 1$ has been discovered in 1998; the expansion is not slowing down under the influence of gravity, but is instead accelerating due to some uniformly distributed, gravitationally repulsive substance accounting for more than 70% of the mass–energy content of the Universe, which is now known as dark energy. Its most common interpretation today is given in terms of the so-called ΛCDM model with a cosmological constant Λ.

This pathbreaking result was obtained almost simultaneously by the Supernova Cosmology Project led by Saul Perlmutter of the Lawrence Berkeley National Laboratory and the University of California at Berkeley, and the High-z Supernova Search Team around Brian Schmidt of the Australian National University in Canberra, and Adam Riess, who is now at the Johns Hopkins University and the Space Telescope Science Institute, both in Baltimore, Maryland. It is presently not clear whether there is any relationship between inflation in the early Universe and the accelerated expansion in the late Universe. Dark energy is somewhat similar to cosmological inflation, but its energy scale of 10^{-12} GeV is about 27 orders of magnitude smaller than the typical energy scale of inflation. In its physical interpretation, the solution for the cosmological constant problem is the clue to further progress in both cosmology and particle physics – in particular, the correct explanation of the smallness of the cosmological constant, and the reason for the approximate equality of its energy density and the matter energy density at the present epoch.

The discovery of the accelerated expansion in the late Universe relied on data from type Ia supernovae. These are fairly reliable standard candles, and can hence be used in the cosmic luminosity-distance determination. The accelerated expansion in the redshift range $z < 1$ has subsequently been and will further be investigated in detail not only through refined data from type Ia supernovae, but also through observations of the temperature fluctuations in the cosmic microwave

background, in particular with the Planck satellite, of baryonic acoustic oscillations, the weak-lensing effect, and through galaxy cluster counts with the South Pole Telescope and other equipment. At larger redshifts, the acceleration becomes a deceleration, owing to the lessening impact of dark energy at earlier times – as has been confirmed by recent supernova data from the Hubble space telescope.

Selected aspects of the vast field of accelerated expansion of our Universe in different epochs are treated in the four selected lectures that are presented in this volume. The first chapter by David Langlois of the Université Paris 7 considers inflation and how it accounts for the primordial seeds of the cosmological perturbations which we can observe today with great precision. It also serves as an introduction to the principles of the Hot Big Bang Model and its limitations. Particular emphasis is placed on the amplification of the quantum vacuum fluctuations during the inflationary phase. The constraints on inflationary models are discussed, as well as more general models of inflation involving multiple fields, and non-Gaussianities of the primordial perturbations.

The second chapter written by Mark Sullivan of Oxford University introduces dark energy in a review of type Ia supernovae results in cosmology. The physics which leads to the near-uniform peak brightness of these supernovae, allowing astronomers to use them for precise luminosity-distance determinations, is explained. Modern SN Ia searches and distance estimation techniques have allowed to measure the average equation of state of dark energy to better than 5% statistical error, when combined with complementary probes of large-scale structure such as baryonic acoustic oscillations. Future prospects for determining dark energy with SN Ia in the next generation of planned experiments are given.

The third chapter by Shinji Tsujikawa of Tokyo University concerns modified gravity models of dark energy. Such theories presently appear to be the most serious competitors to conventional dark-energy models based on a cosmological constant, or its time-dependent counterparts arising from a scalar field, although the physical origin of the modifications of gravity are not always clear. A number of modified-gravity models that satisfy local gravity and cosmological constraints are presented. Signatures that may distinguish modified gravity models from ΛCDM cosmology are discussed. The braneworld models treated in this chapter as possible candidates for late-time cosmic acceleration are, however, ruled out from observational constraints.

Finally, the last lecture by Licia Verde from the Universitat de Barcelona about statistical methods in cosmology gives a summary of the currently available statistical methods that are indispensable for the analysis of cosmological data, and are thus necessary prerequisites for many of the results that have been presented in this volume and eleswhere in cosmological research.

The lectures have grown out of the annual Winter School of the Transregional Research Center TRR33 “The Dark Universe” of the joint Universities Heidelberg, Bonn, and LMU Munich. The center is funded by the Deutsche Forschungsgemeinschaft. The first Winter School was established in 2007 by the Research Center following the request and initiative of the young researchers, including postdoctoral, doctoral, and master students. The most active group of them acted as organizers and

set up the school in the Italian mountain resort, Tonale, at moderate cost but great scientific benefit for the center. The main idea was to present "theory for observers and observations for theorists," to initiate discussions and joint projects between theorists, observers, and scientists working with simulation methods.

The school was a big success, and it is scheduled to continue every year during the funding period of TRR33. The four selected lectures in this book arose from the second school in 2008, which was mainly organized by the young researchers, M. Baldi, C. Byrnes, T. Koivisto, M. Maturi, C. Mignone, D. Mota, V. Pettorino, G. Robbers, M. Viola, and J.C. Waizmann, with the help of some more senior people in the background. The four authors of this Lecture Notes volume have expressed their gratitude to the organizers for setting up this very useful and enjoyable Research School and for their hospitality. We all hope that this endeavor will contribute further to the bright future of dark energy.

Heidelberg Georg Wolschin
September 2009

Contents

Inflation and Cosmological Perturbations

D. Langlois

Summary The purpose of these lectures is to give a pedagogical introduction to inflation and the production of the primordial perturbations, as well as a review of some of the latest developments in this domain.

After a short introduction, we review the main principles of the Hot Big Bang model, as well as its limitations. These deficiencies provide the motivation for the study of a cosmological phase of accelerated expansion, called inflation, which can be induced by a slow-rolling scalar field. A few illustrative models are presented. We then turn to the analysis of cosmological perturbations and explain how the vacuum quantum fluctuations are amplified during an inflationary phase. The next step consists in relating the perturbations generated during inflation to the perturbations of the cosmological fluid in the standard radiation-dominated phase. One can thus confront the predictions of inflationary models with cosmological observations, such as the measurements of the Cosmic Microwave Background or the large-scale structure surveys. The present constraints on inflationary models are discussed.

The final part of these lectures gives a review of more general models of inflation, involving multiple fields or non-standard kinetic terms. Although more complicated, these models are usually motivated by high-energy physics and they can lead to specific signatures that are not expected in the simplest models of inflation. After introducing a very general formalism to describe perturbations in multi-field models with arbitrary kinetic terms, several interesting cases are presented. We also stress the role of entropy perturbations in the context of multi-field models. Finally, we discuss in detail the non-Gaussianities of the primordial perturbations and some models that could produce a detectable level of non-Gaussianities.

D. Langlois (✉)
APC (CNRS-Université Paris 7), 10, rue Alice Domon et Léonie Duquet, 75205 Paris Cedex 13, France
e-mail: langlois@apc.univ-paris7.fr

Langlois, D.: *Inflation and Cosmological Perturbations*. Lect. Notes Phys. **800**, 1–57 (2010)
DOI 10.1007/978-3-642-10598-2_1

1 Introduction

Inflation is today the main theoretical framework that describes the early Universe and that can account for the present observational data. In 30 years of existence, inflation has survived, in contrast with earlier competitors, the tremendous improvement of cosmological data. In particular, the fluctuations of the Cosmic Microwave Background (CMB) had not yet been measured when inflation was invented, whereas they give us today a remarkable picture of the cosmological perturbations in the early Universe. In the future, one can hope that more precise measurements of the primordial cosmological perturbations will allow us to go one step further in the confrontation of inflation models with data, and especially to discriminate between the many different possible realizations of inflation.

The purpose of these lectures is 2-fold. The first goal is to explain, in a simple way and starting from first principles as much as possible, the conceptual basis of inflation and the elementary steps to calculate the cosmological perturbations predicted by the simplest models. The second objective of these lectures is to give an overview of the latest developments on inflation, in particular the study of more general models of inflation involving several scalar fields or non-standard kinetic terms. Although more complicated, these models can give very specific signatures in the primordial cosmological perturbations, in particular non-Gaussianities and isocurvature perturbations.

There is a huge literature on inflation and these lectures cover only a few topics, with a list of references that is far from exhaustive. More details and more references can be found in several textbooks (see, e.g. [1–3]) and many reviews (including for instance, [4–8]; more specialized reviews will also be mentioned in the text). A novel feature of the present lectures is to introduce the most modern approach to the computation of perturbations. This has the advantage to be easily applicable to the study of non-linear perturbations, which has recently become an extremely active topic.

The outline of these lectures is the following. In the next section, we recall the basic elements of the Hot Big Bang model and discuss its limitations, which motivate inflation. Homogeneous inflation is introduced in Sect. 3. In Sect. 4, we turn to the theory of linear cosmological perturbations and explain how they are generated during an inflationary phase. The following section, Sect. 5, is devoted to the link between primordial perturbations and present cosmology, and thus to the confrontation of inflation models with the data. In Sect. 6, more general models of inflation are considered, with a discussion of several specific scenarios, which have attracted a lot of attention recently. Section 7 is devoted to the primordial non-Gaussianities. And we conclude in the last section.

2 The Hot Big Bang Model

Modern cosmology is based on the theory of general relativity, according to which our Universe is described by a four-dimensional geometry $g_{\mu\nu}$ that satisfies Einstein's equation

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2} R\, g_{\mu\nu} = 8\pi G\, T_{\mu\nu}, \tag{1}$$

where $R_{\mu\nu}$ is the Ricci tensor, $R \equiv g^{\mu\nu} R_{\mu\nu}$ the scalar curvature and $T_{\mu\nu}$ the energy–momentum tensor that describes the matter distribution.

2.1 The Friedmann Equations

One of the main assumptions of cosmology, which has been confirmed by observations so far, is to consider, as a first approximation, the Universe as being homogeneous and isotropic. Note that these symmetries define implicitly a particular "slicing" of spacetime, in which the space-like hypersurfaces are homogeneous and isotropic. A different slicing of the *same* spacetime would give space-like hypersurfaces that are not homogeneous and isotropic.

Homogeneity and isotropy turn out to be very restrictive and the only geometries compatible with these requirements are the FLRW (Friedmann–Lemaître–Robertson–Walker) spacetimes, with metric

$$ds^2 = -dt^2 + a^2(t) \left[\frac{dr^2}{1 - \kappa r^2} + r^2 \left(d\theta^2 + \sin^2 \theta d\phi^2 \right) \right], \tag{2}$$

where $\kappa = 0, -1, 1$ determines the curvature of spatial hypersurfaces, respectively, flat, elliptic or hyperbolic. Moreover, the matter content compatible with homogeneity and isotropy is necessarily characterized by an energy–momentum tensor of the form

$$T^{\mu}_{\ \nu} = \mathrm{Diag}\left(-\rho(t), P(t), P(t), P(t)\right), \tag{3}$$

where ρ corresponds to an energy density and P to a pressure.

Substituting the metric (2) and the energy–momentum tensor (3) into Einstein's equations (1) gives the Friedmann equations,

$$\left(\frac{\dot{a}}{a} \right)^2 = \frac{8\pi G \rho}{3} - \frac{\kappa}{a^2}, \tag{4}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3} \left(\rho + 3P \right), \tag{5}$$

which govern the time evolution of the scale factor $a(t)$.

An immediate consequence of the two above equations is the *continuity equation*

$$\dot{\rho} + 3H \left(\rho + P \right) = 0, \tag{6}$$

where $H \equiv \dot{a}/a$ is the *Hubble parameter*. The continuity equation can also be obtained directly from the energy–momentum conservation

$$\nabla_\mu T^\mu_{\ \nu} = 0, \tag{7}$$

where ∇ denotes the covariant derivative associated with the metric $g_{\mu\nu}$.

The cosmological evolution can be determined once the equation of state for the matter is specified. Let us assume here $P = w\rho$ with w constant, which includes the two main types of matter that play an important rôle in cosmology, namely non-relativistic matter ($w \simeq 0$) and a gas of relativistic particles ($w = 1/3$). The conservation equation (6) can be integrated to give

$$\rho \propto a^{-3(1+w)}. \tag{8}$$

Substituting in (4), one finds, for $\kappa = 0$,

$$a(t) \propto t^{\frac{2}{3(1+w)}}, \tag{9}$$

which thus gives the evolution $a(t) \propto t^{1/2}$ for relativistic matter and $a(t) \propto t^{2/3}$ for non-relativistic matter. Note that a different cosmological evolution, governed by modified Friedmann's equations, can be envisaged in the primordial Universe, as for example, in the context of brane cosmology (see, e.g. [9]), but this possibility will not be discussed in these notes.

The present cosmological observations seem to indicate that our Universe is currently accelerating. The simplest way to account for this acceleration is to assume the presence of a *cosmological constant* Λ in Einstein's equations, i.e. an additional term $\Lambda g_{\mu\nu}$ on the left-hand side of (1). By moving this term on the right-hand side of Einstein's equations it can also be interpreted as an energy–momentum tensor with equation of state $P = -\rho$, where ρ is time independent. This leads, for $\kappa = 0$ and without any other matter, to an exponential evolution of the scale factor

$$a(t) \propto \exp(Ht). \tag{10}$$

In our Universe, several species with different equations of state coexist, and it has become customary to characterize their relative contributions by the dimensionless parameters

$$\Omega_{(i)} \equiv \frac{8\pi G\rho_0^{(i)}}{3H_0^2}, \tag{11}$$

where $\rho_0^{(i)}$ denote the present energy densities of the various species, and H_0 is the present Hubble parameter. The first Friedmann equation (4), evaluated at the present time, implies

$$\Omega_0 = \sum_i \Omega_{(i)} = 1 + \frac{\kappa}{a_0^2 H_0^2}. \tag{12}$$

One can infer from present observations the following parameters: $\Omega_{\rm m} \simeq 0.3$ for non-relativistic matter (which includes a small baryonic component $\Omega_{\rm b} \simeq 0.05$), $\Omega_\Lambda \simeq 0.7$ for a "dark energy" component (compatible with a cosmological constant), $\Omega_\gamma \simeq 5 \times 10^{-5}$ for the photons, and a total Ω_0 close to 1, i.e. no detectable deviation from flatness.

2.2 The Shortcomings of the Standard Big Bang Model

The standard Big Bang model has encountered remarkable successes, in particular with primordial nucleosynthesis and the CMB, and it remains today a cornerstone in our understanding of the present and past Universe. However, a few intriguing facts remain unexplained in the strict scenario of the Hot Big Bang model and seem to necessitate a larger framework. We review now the main problems:

- *Homogeneity problem:* A first question is why the approximation of homogeneity and isotropy turns out to be so good. Indeed, inhomogeneities are unstable, because of gravitation, and they tend to grow with time. It can be verified, for instance, with the CMB that inhomogeneities were much smaller at the last scattering epoch than today. One thus expects that these homogeneities were still smaller further back in time. How to explain a Universe so smooth in its past ?
- *Flatness problem:* Another puzzle lies in the (spatial) flatness of our Universe. Indeed, the first Friedmann equation (4) implies

$$\Omega - 1 \equiv \frac{8\pi G\rho}{3H^2} - 1 = \frac{\kappa}{a^2H^2}. \tag{13}$$

  In standard cosmology, the scale factor behaves like $a \sim t^p$ with $p < 1$ ($p = 1/2$ for radiation and $p = 2/3$ for non-relativistic matter). As a consequence, $(aH)^{-2}$ grows with time and $|\Omega - 1|$ must thus diverge with time. Therefore, in the context of the standard model, the quasi-flatness observed today requires an extreme fine-tuning of Ω near 1 in the early Universe.
- *Horizon problem:* One of the most fundamental problems in standard cosmology is certainly the *horizon problem*. The (particle) *horizon* is the maximal distance that can be covered by a light ray. For a light-like radial trajectory $dr = a(t)dt$, and the horizon is thus given by

$$d_H(t) = a(t) \int_{t_i}^{t} \frac{dt'}{a(t')} = a(t) \frac{t^{1-q} - t_i^{1-q}}{1-q}, \tag{14}$$

  where the last equality is obtained by assuming $a(t) \sim t^q$ and t_i is some initial time.

  In standard cosmology ($q < 1$), the integral converges in the limit $t_i = 0$ and the horizon has a finite size, of the order of the so-called Hubble radius H^{-1}:

$$d_H(t) = \frac{q}{1-q} H^{-1}. \tag{15}$$

It is also useful to consider the *comoving Hubble radius*, $(aH)^{-1}$, which represents the fraction of comoving space in causal contact. One finds that it *grows* with time, which means that the *fraction of the Universe in causal contact increases with time* in the context of standard cosmology. But the CMB tells us that the Universe was quasi-homogeneous at the time of last scattering on a scale encompassing many regions a priori causally independent. How to explain this ?

A solution to the horizon problem and to the other puzzles is provided by the inflationary scenario, which we will examine in the next section. The basic idea is to "decouple" the causal size from the Hubble radius, so that the real size of the horizon region in the standard radiation-dominated era is much larger than the Hubble radius. Such a situation occurs if the comoving Hubble radius *decreases* sufficiently in the very early Universe. The corresponding condition is

$$\ddot{a} > 0, \tag{16}$$

i.e. the Universe undergoes a *phase of acceleration*.

3 Inflation

The broadest definition of inflation is that it corresponds to a phase of acceleration of the Universe,

$$\ddot{a} > 0. \tag{17}$$

In this sense, the current cosmological observations, if correctly interpreted, mean that our present Universe is undergoing an inflationary phase. It is worth noting that many of the models suggested for inflation have been adapted to account for the present acceleration. We are, however, interested here in an inflationary phase taking place in the *early* Universe, thus characterized by very different energy scales. Another difference is that inflation in the early Universe *must end* to leave room to the standard radiation-dominated cosmological phase.

Cosmological acceleration requires, according to the second Friedmann equation (5), an equation of state satisfying

$$P < -\frac{1}{3}\rho, \tag{18}$$

condition which looks at first view rather exotic.

A very simple example giving such an equation of state is a cosmological constant, corresponding to a cosmological fluid with the equation of state

$$P = -\rho. \tag{19}$$

However, a strict cosmological constant leads to exponential inflation *forever* which cannot be followed by a radiation era. Another possibility is a scalar field, which we now discuss in some details.

3.1 Cosmological Scalar Fields

The dynamics of a scalar field minimally coupled to gravity is governed by the action

$$S_\phi = \int d^4x\sqrt{-g}\left(-\frac{1}{2}\partial^\mu\phi\partial_\mu\phi - V(\phi)\right), \tag{20}$$

where $g \equiv \det(g_{\mu\nu})$ and $V(\phi)$ is the potential of the scalar field. The corresponding energy–momentum tensor, obtained by varying the action (20) with respect to the metric, is given by

$$T_{\mu\nu} = \partial_\mu\phi\partial_\nu\phi - g_{\mu\nu}\left(\frac{1}{2}\partial^\sigma\phi\partial_\sigma\phi + V(\phi)\right). \tag{21}$$

In the homogeneous and isotropic geometry (2), the energy–momentum tensor is of the perfect fluid form, with the energy density

$$\rho = -T^0_0 = \frac{1}{2}\dot{\phi}^2 + V(\phi), \tag{22}$$

where one recognizes the sum of a kinetic energy and potential energy and the pressure

$$P = \frac{1}{2}\dot{\phi}^2 - V(\phi). \tag{23}$$

The equation of motion for the scalar field is the Klein–Gordon equation, obtained by taking the variation of the above action (20) with respect to the scalar field,

$$\nabla^\mu\nabla_\mu\phi = \frac{dV}{d\phi}, \tag{24}$$

which reduces to

$$\ddot{\phi} + 3H\dot{\phi} + V' = 0 \tag{25}$$

in a homogeneous and isotropic Universe.

The system of equations governing the dynamics of the scalar field and of the cosmological geometry is thus given by

$$H^2 = \frac{8\pi G}{3}\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right), \tag{26}$$

$$\ddot{\phi} + 3H\dot{\phi} + V' = 0, \tag{27}$$

$$\dot{H} = -4\pi G\dot{\phi}^2. \tag{28}$$

The last equation can be derived from the first two and is therefore redundant.

3.2 The Slow-Roll Regime

The dynamical system (26), (27), and (28) does not always give an accelerated expansion but it does so in the so-called *slow-roll regime* when the potential energy of the scalar field dominates over its kinetic energy.

More specifically, the slow-roll approximation consists in neglecting the kinetic energy of the scalar field , $\dot{\phi}^2$ in (26) and the acceleration $\ddot{\phi}$ in the Klein–Gordon equation (27). One then gets the simplified system

$$H^2 \simeq \frac{8\pi G}{3}V, \tag{29}$$

$$3H\dot{\phi} + V' \simeq 0. \tag{30}$$

Let us now examine in which regime this approximation is valid. From (30), the velocity of the scalar field is given by

$$\dot{\phi} \simeq -\frac{V'}{3H}. \tag{31}$$

Substituting this relation in the condition $\dot{\phi}^2/2 \ll V$ yields the requirement:

$$\varepsilon_V \equiv \frac{M_P^2}{2}\left(\frac{V'}{V}\right)^2 \ll 1, \tag{32}$$

where we have introduced the *reduced Planck mass*

$$M_P \equiv \frac{1}{\sqrt{8\pi G}}. \tag{33}$$

Alternatively, one can use the parameter

$$\varepsilon \equiv -\frac{\dot{H}}{H^2}, \tag{34}$$

which coincides with ε_V at leading order in slow-roll, since $\varepsilon = \dot{\phi}^2/(2M_P^2H^2)$.

Similarly, $\ddot{\phi} \ll V'$ implies, after using the time derivative of (31) and (29), the condition

$$\eta_V \equiv M_P^2 \frac{V''}{V} \ll 1. \tag{35}$$

In summary, the slow-roll approximation is valid when the conditions ε_V,$\eta_V \ll 1$ are satisfied by the potential, which means that the slope and the curvature of the potential, in Planck units, must be sufficiently small.

3.3 Number of e-Folds

Inflation must last long enough, in order to solve the problems of the Hot Big Bang model. To investigate this question, one usually introduces the *number of e-folds before the end of inflation*, denoted by N, and simply defined by

$$N = \ln \frac{a_{\rm end}}{a}, \tag{36}$$

where $a_{\rm end}$ is the value of the scale factor at the end of inflation and a is a fiducial value for the scale factor during inflation. By definition, N *decreases* during the inflationary phase and reaches zero at its end.

In the slow-roll approximation, it is possible to express N as a function of the scalar field. Since $dN = -d\ln a = -Hdt = -(H/\dot{\phi})d\phi$, one easily finds, using (31) and (29), that

$$N(\phi) \simeq \int_{\phi}^{\phi_{\rm end}} \frac{V}{M_P^2 V'} d\phi. \tag{37}$$

Given an explicit potential $V(\phi)$, one can in principle integrate the above expression to obtain N in terms of ϕ. This will be illustrated below for some specific models.

Let us now discuss the link between N and the present cosmological scales. Let us consider a given scale characterized by its comoving wavenumber $k = 2\pi/\lambda$. This scale crossed out the Hubble radius, during inflation, at an instant $t_*(k)$ defined by

$$k = a(t_*)H(t_*). \tag{38}$$

To get a rough estimate of the number of e-foldings of inflation that are needed to solve the horizon problem, let us first ignore the transition from a radiation era to a matter era and assume for simplicity that the inflationary phase was followed instantaneously by a radiation phase that has lasted until now. During the radiation phase, the comoving Hubble radius $(aH)^{-1}$ increases like a. In order to solve the horizon problem, the increase of the comoving Hubble radius during the standard evolution must be compensated by *at least* a decrease of the same amount during inflation. Since the comoving Hubble radius roughly scales like a^{-1} during inflation,

the minimum amount of inflation is simply given by the number of e-folds between the end of inflation and today

$$\ln\left(a_0/a_{\rm end}\right) = \ln\left(T_{\rm end}/T_0\right) \sim \ln\left(10^{29}(T_{\rm end}/10^{16}\,{\rm GeV})\right), \tag{39}$$

i.e. around 60 e-folds for a temperature $T \sim 10^{16}\,{\rm GeV}$ at the beginning of the radiation era. As we will see later, this energy scale is typical of inflation in the simplest models (Fig. 1).

This determines roughly the number of e-folds $N(k_0)$ between the moment when the scale corresponding to our present Hubble radius $k_0 = a_0 H_0$ exited the Hubble radius during inflation and the end of inflation. The other lengthscales of cosmological interest are *smaller* than k_0^{-1} and therefore exited the Hubble radius during inflation *after* the scale k_0, whereas they entered the Hubble radius during the standard cosmological phase (either in the radiation era for the smaller scales or in the matter era for the larger scales) *before* the scale k_0.

A more detailed calculation, which distinguishes between the energy scales at the end of inflation and after the reheating, gives for the number of e-folds between the exit of the mode k and the end of inflation (see, e.g. [2, 10])

$$N(k) \simeq 62 - \ln\frac{k}{a_0 H_0} + \ln\frac{V_k^{1/4}}{10^{16}{\rm GeV}} + \ln\frac{V_k^{1/4}}{V_{\rm end}^{1/4}} + \frac{1}{3}\ln\frac{\rho_{\rm reh}^{1/4}}{V_{\rm end}^{1/4}}. \tag{40}$$

Since the smallest scale of cosmological relevance is of the order of 1 Mpc, the range of cosmological scales covers about 9 e-folds.

The above number of e-folds is altered if one changes the thermal history of the Universe between inflation and the present time by including, for instance, a period of so-called thermal inflation.

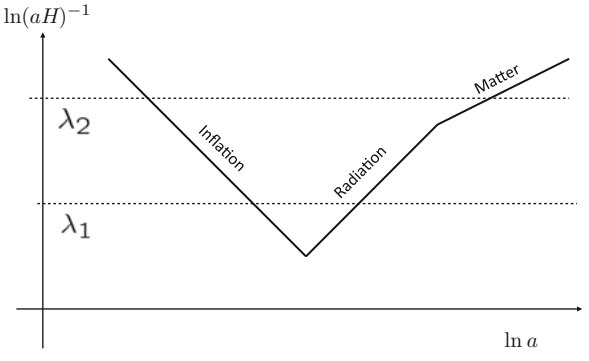

Fig. 1 Evolution of the comoving Hubble radius $\lambda_H = (aH)^{-1}$, during inflation, radiation-dominated era and matter-dominated era. The *horizontal dashed lines* correspond to two different comoving lengthscales: the larger scales cross out the Hubble radius *earlier* during inflation and reenter the Hubble radius *later* in the standard cosmological era

3.4 A Few Examples

It is now time to illustrate all the points discussed above with some specific potentials.

3.4.1 Power Law Potential

We consider first the case of power law monomial potentials, of the form

$$V(\phi) = \lambda \phi^p, \tag{41}$$

which have been abundantly studied in the literature. In particular, the above potentials include the case of a free massive scalar field, $V(\phi) = m^2\phi/2$.

The slow-roll parameters are given by

$$\varepsilon = \frac{p^2 M_P}{2\phi^2}, \qquad \eta = p(p-1)\frac{M_P^2}{\phi^2}. \tag{42}$$

The slow-roll conditions $\varepsilon \ll 1$ and $\eta \ll 1$ thus imply

$$\phi \gg p\, M_P, \tag{43}$$

which means that the scalar field amplitude must be well above the Planck mass during inflation.

After substituting the potential (41) into the slow-roll equations of motion (29) and (30), one can integrate them explicitly to get

$$\phi^{2-p/2} - \phi_i^{2-p/2} = -\frac{2p}{4-p}\sqrt{\frac{\lambda}{3}}\, M_P\, (t - t_i) \tag{44}$$

for $p \neq 4$ and

$$\phi = \phi_i \exp\left[-4\sqrt{\frac{\lambda}{3}}\, M_P (t - t_i)\right] \tag{45}$$

for $p = 4$.

One can also express the scale factor as a function of the scalar field [and thus as a function of time by substituting the above expression for $\phi(t)$] by using $d\ln a/d\phi = H/\dot{\phi} \simeq -\phi/(pM_P^2)$. One finds

$$a = a_{\rm end} \exp\left[-\frac{\left(\phi^2 - \phi_{\rm end}^2\right)}{2p\, M_P^2}\right]. \tag{46}$$

Defining the end of inflation by $\varepsilon = 1$, one gets $\phi_{\rm end} = p\,M_P/\sqrt{2}$ and the number of e-folds is thus given by

$$N(\phi) \simeq \frac{\phi^2}{2pM_P^2} - \frac{p}{4}. \tag{47}$$

This can be inverted, so that

$$\phi(N) \simeq \sqrt{2Np}\,M_P, \tag{48}$$

where we have ignored the second term of the right-hand side of (47), consistently with the condition (43).

3.4.2 Exponential Potential

Cosmological scalar fields with a potential of the form

$$V = V_0 \exp\left(-\sqrt{\frac{2}{q}}\frac{\phi}{M_P}\right) \tag{49}$$

admit an *exact* solution (i.e. valid beyond the slow-roll approximation) of the system (26), (27), and (28), with a power law scale factor, i.e.

$$a(t) \propto t^q. \tag{50}$$

The evolution of the scalar field is given by the expression

$$\phi(t) = \sqrt{2q}\,M_P \ln\left[\sqrt{\frac{V_0}{q(3q-1)}}\frac{t}{M_P}\right]. \tag{51}$$

Note that one recovers the slow-roll approximation in the limit $q \gg 1$, since the slow-roll parameters are given by

$$\varepsilon_V = \frac{1}{q} \qquad \eta_V = \frac{2}{q}. \tag{52}$$

3.4.3 Hybrid Inflation

In this type of model, the potential contains a constant piece in addition to a power law potential. The simplest example is

$$V(\phi) = V_0 + \frac{1}{2}m^2\phi^2. \tag{53}$$

In fact, the full model relies on the presence of two scalar fields, where one plays the traditional role of the inflaton, while the other is necessary to end inflation. In the original model of hybrid inflation [11], one starts from the potential

$$V(\phi,\psi) = \frac{1}{2}m^2\phi^2 + \frac{1}{2}\lambda'\psi^2\phi^2 + \frac{1}{4}\lambda\left(M^2 - \psi^2\right)^2. \tag{54}$$

For values of the field ϕ larger than the critical value $\phi_c = \lambda M^2/\lambda'$, the potential for ψ has its minimum at $\psi = 0$. This is the case during inflation. ψ is thus trapped in this minimum $\psi = 0$, so that the effective potential for the scalar field ϕ, which plays the role of the inflaton, is given by (53) with $V_0 = \lambda M^4/4$. During the inflationary phase, the field ϕ slow-rolls until it reaches the critical value ϕ_c. The shape of the potential for ψ is then modified and new minima appear in $\psi = \pm M$. ψ will thus roll down into one of these new minima and, as a consequence, inflation will end.

During the inflationary phase, the slow-roll parameters are given by

$$\varepsilon = \frac{m^2 M_P^2 \tilde{\phi}^2}{V_0(1+\tilde{\phi}^2)^2}, \qquad \eta = \frac{m^2 M_P^2}{V_0(1+\tilde{\phi}^2)}, \tag{55}$$

where we have introduced the rescaled scalar field $\tilde{\phi}$, which is dimensionless and defined so that $V = V_0(1+\tilde{\phi}^2)$. Note that there are two limiting regimes: if $\tilde{\phi} \gg 1$, the constant term is negligible and one recovers a power law potential with $p = 2$; if $\tilde{\phi} \ll 1$, V_0 dominates and the potential is extremely flat with $\varepsilon \ll \eta$.

3.5 The Inflationary "Zoology"

3.5.1 Historical Perspective

The first model of inflation is usually traced back to Alan Guth [12] in 1981, although one can see the model of Alexei Starobinsky [13] as a precursor. Guth's model, which is today named as *old inflation* is based on a first-order phase transition, from a false vacuum with non-zero energy, which generates an exponential inflationary phase, into a true vacuum with zero energy density. The true vacuum phase appears in the shape of bubbles via quantum tunneling. The problem with this inflationary model is that, in order to get sufficient inflation to solve the problems of the standard model mentioned earlier, the nucleation rate must be sufficiently small; but, then, the bubbles never coalesce because the space that separates the bubbles undergoes inflation and expands too rapidly. Therefore, the first model of inflation is not phenomenologically viable.

After this first and unsuccessful attempt, a new generation of inflationary models appeared, usually denoted as *new inflation* models [14, 15]. They rely on a second-order phase transition, based on thermal corrections of the effective potential and thus assume that the scalar field is in thermal equilibrium.

This hypothesis of thermal equilibrium was given up in the third generation of models, initiated by Andrei Linde, and whose generic name is *chaotic inflation* [16]. This allows to use extremely simple potentials, quadratic or quartic, which lead to inflationary phases when the scalar field is displaced from the origin with values of the order of several Planck masses.

In the last few years, there has been an intense activity in building inflationary models based on high-energy theories, in particular in the context of supersymmetry and string theory. Details can be found in several recent reviews[17–21].

3.5.2 Classification

There exist a huge number of models of inflation. As far as *single-field* models are concerned,[1] it is convenient to regroup them into three broad categories (Fig. 2):

- Large field models ($0 < \eta \leq \varepsilon$)
  The scalar field is displaced from its stable minimum by $\Delta\phi \sim M_P$. This includes the chaotic models with monomial potentials

$$V(\phi) = \Lambda^4 \left(\frac{\phi}{\mu}\right)^p , \tag{56}$$

  or the exponential potential

$$V(\phi) = \Lambda^4 \exp\left(\phi/\mu\right), \tag{57}$$

  which have already been discussed.
- Small field models ($\eta < 0 < \varepsilon$)
  In this type of models, the scalar field is rolling away from an unstable maximum of the potential. This is a characteristic feature of spontaneous symmetry breaking. A typical potential is

$$V(\phi) = \Lambda^4 \left[1 - \left(\frac{\phi}{\mu}\right)^p\right], \tag{58}$$

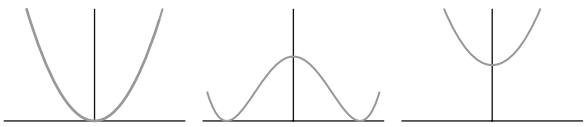

Fig. 2 Schematic potential for the three main categories of inflationary models: large-field models, small-field models, hybrid models

[1] Or at least *effectively* single field during inflation (the hybrid models require a second field to *end* inflation as discussed earlier).

which can be interpreted as the lowest-order term in a Taylor expansion about the origin. Historically, this potential shape appeared in the so-called 'new inflation' scenario.

A particular feature of these models is that tensor modes are much more suppressed with respect to scalar modes than in the large-field models, as it will be shown later.

- Hybrid models ($0 < \varepsilon < \eta$)
  Although a second scalar field is needed to end inflation, hybrid models correspond effectively to single-field models with a potential characterized by $V''(\phi) > 0$ and $0 < \varepsilon < \eta$. A typical potential is

$$V(\phi) = \Lambda^4 \left[1 + \left(\frac{\phi}{\mu}\right)^p\right]. \tag{59}$$

  Once more, this potential can be seen as the lowest order in a Taylor expansion about the origin.

  In the case of hybrid models, the value ϕ_N of the scalar field as a function of the number of e-folds before the end of inflation is not determined by the above potential and, therefore, (ϕ_N/μ) can be considered as a freely adjustable parameter.

4 Quantum Fluctuations and "Birth" of Cosmological Perturbations

So far, we have concentrated our attention on strictly homogeneous and isotropic aspects of cosmology. Of course, this idealized version, although extremely useful, is not sufficient to account for real cosmology and it is now time to turn to the study of deviations from homogeneity and isotropy.

In cosmology, inhomogeneities grow because of the attractive nature of gravity, which implies that inhomogeneities were much smaller in the past. As a consequence, for most of their evolution, inhomogeneities can be treated as *linear perturbations*. The linear treatment ceases to be valid on small scales in our recent past, hence the difficulty to reconstruct the primordial inhomogeneities from large-scale structure, but it is quite adequate to describe the fluctuations of the CMB at the time of last scattering. This is the reason why the CMB is currently the best observational probe of primordial inhomogeneities.

In this section, we concentrate on the perturbations of the inflaton and show how the accelerated expansion during inflation converts its *initial vacuum quantum fluctuations* into "macroscopic" cosmological perturbations (see [22–27] for some of the historical works). In this sense, inflation provides us with "natural" initial conditions. We will also see how the perturbations of the inflaton can be translated into perturbations of the geometry.

4.1 Massless Scalar Field in de Sitter

As a warming-up, it is instructive to discuss the case of a massless scalar field in a so-called de Sitter Universe, or a cosmological spacetime with exponential expansion, $a \propto \exp(Ht)$,

$$ds^2 = -dt^2 + a^2(t)d\mathbf{x}^2, \quad a(t) = e^{Ht}. \tag{60}$$

It turns out it is more convenient to use, instead of the cosmic time t, a conformal time τ defined by

$$\tau = \int \frac{dt}{a(t)}, \tag{61}$$

so that the metric takes the particularly simple form

$$ds^2 = a^2(\tau)\left[-d\tau^2 + d\mathbf{x}^2\right]. \tag{62}$$

In the de Sitter case, the conformal time is given by

$$\tau = -\frac{e^{-Ht}}{H} = -\frac{1}{aH}, \tag{63}$$

so that the scale factor in terms of τ is simply

$$a(\tau) = -\frac{1}{H\tau}. \tag{64}$$

The conformal time is here negative (so that the scale factor is positive) and goes from $-\infty$ to 0.

The action for a massless scalar field is given by

$$S = \int d^4x\sqrt{-g}\left(-\frac{1}{2}\partial_\mu\phi\partial^\mu\phi\right) = \int d\tau\ d^3x\ a^4\left[\frac{1}{2a^2}\phi'^2 - \frac{1}{2a^2}\nabla\phi^2\right], \tag{65}$$

where we have substituted in the action the cosmological metric (62) and where a prime denotes a derivative with respect to the conformal time τ. Note that, whereas we still allow for spatial variations of the scalar field, i.e. inhomogeneities, we will assume here, somewhat inconsistently, that the geometry is completely fixed as homogeneous. We will deal later with the question of the metric perturbations.

It is possible to eliminate the factor a^2 in front of the kinetic term ϕ'^2 by introducing the new function

$$u = a\,\phi. \tag{66}$$

This will generate a term proportional to uu' but one can get rid of it by an integration by parts. The action (65) can then be rewritten as

$$S = \frac{1}{2}\int d\tau\ d^3x \left[u'^2 - \nabla u^2 + \frac{a''}{a}u^2\right]. \tag{67}$$

The first two terms are familiar since they are the same as in the action for a free massless scalar field in Minkowski spacetime. The fact that our scalar field here lives in de Sitter spacetime rather than Minkowski has been reexpressed as a *time-dependent effective mass*

$$m_{\rm eff}^2 = -\frac{a''}{a} = -\frac{2}{\tau^2}. \tag{68}$$

Let us now proceed to the quantization of the scalar field u by using the standard procedure of quantum field theory. One first turns u into a quantum field denoted by $\hat{u}$, which we expand in Fourier space as

$$\hat{u}(\tau,\mathbf{x}) = \frac{1}{(2\pi)^{3/2}}\int d^3k \left\{\hat{a}_{\mathbf{k}} u_k(\tau) e^{i\mathbf{k}.\mathbf{x}} + \hat{a}_{\vec{k}}^{\dagger} u_k^*(\tau) e^{-i\mathbf{k}.\mathbf{x}}\right\}, \tag{69}$$

where the $\hat{a}^{\dagger}$ and $\hat{a}$ are creation and annihilation operators, respectively, satisfying the usual commutation rules

$$\left[\hat{a}_{\mathbf{k}},\hat{a}_{\mathbf{k}'}\right] = \left[\hat{a}_{\mathbf{k}}^{\dagger},\hat{a}_{\mathbf{k}'}^{\dagger}\right] = 0, \quad \left[\hat{a}_{\mathbf{k}},\hat{a}_{\mathbf{k}'}^{\dagger}\right] = \delta(\mathbf{k}-\mathbf{k}'). \tag{70}$$

The function $u_k(\tau)$ is a complex time-dependent function that must satisfy the *classical* equation of motion in Fourier space, namely

$$u_k'' + \left(k^2 - \frac{a''}{a}\right)u_k = 0, \tag{71}$$

which is simply the equation of motion for an oscillator with a time-dependent mass. In the case of a massless scalar field in Minkowski spacetime, this effective mass is zero ($a''/a = 0$) and one usually takes

$$u_k = \sqrt{\frac{\hbar}{2k}}e^{-ik\tau} \quad \text{(Minkowski)}, \tag{72}$$

where the choice for the normalization factor will be clear below. In the case of de Sitter, one can solve explicitly the above equation with $a''/a = 2/\tau^2$ and the general solution is given by

$$u_k = \alpha e^{-ik\tau}\left(1 - \frac{i}{k\tau}\right) + \beta e^{ik\tau}\left(1 + \frac{i}{k\tau}\right). \tag{73}$$

Canonical quantization consists in imposing the following commutation rules on the τ =constant hypersurfaces:

$$\left[\hat{u}(\tau,\mathbf{x}),\hat{u}(\tau,\mathbf{x}')\right] = \left[\hat{\pi}_u(\tau,\mathbf{x}),\hat{\pi}_u(\tau,\mathbf{x}')\right] = 0 \tag{74}$$

and

$$\left[\hat{u}(\tau,\mathbf{x}),\hat{\pi}_u(\tau,\mathbf{x}')\right] = i\hbar\delta(\mathbf{x}-\mathbf{x}'), \tag{75}$$

where $\pi_u \equiv \delta S/\delta u'$ is the conjugate momentum of u. In the present case, $\pi_u = u'$ since the kinetic term is canonical.

Subtituting the expansion (69) in the commutator (75), and using the commutation rules for the creation and annihilation operators (70), one obtains the relation

$$u_k u_k'^* - u_k^* u_k' = i\hbar, \tag{76}$$

which determines the normalization of the Wronskian.

The choice of a specific function $u_k(\tau)$ corresponds to a particular prescription for the physical vacuum $|0\rangle$, defined by

$$\hat{a}_{\mathbf{k}}|0\rangle = 0. \tag{77}$$

A different choice for $u_k(\tau)$ is associated to a different decomposition into creation and annihilation modes and thus to a different vacuum.

Let us now note that the wavelength associated with a given mode k can always be found *within* the Hubble radius provided one goes sufficiently far backwards in time, since the comoving Hubble radius is shrinking during inflation. In other words, for $|\tau|$ sufficiently big, one gets $k|\tau| \gg 1$. Moreover, for a wavelength smaller than the Hubble radius, one can neglect the influence of the curvature of spacetime and the mode behaves as in a Minkowski spacetime, as can also be checked explicitly with the equation of motion (71) (the effective mass is negligible for $k|\tau| \gg 1$). Therefore, the most natural physical prescription is to take the particular solution that corresponds to the usual Minkowski vacuum, i.e. $u_k \sim \exp(-ik\tau)$, in the limit $k|\tau| \gg 1$. In view of (73), this corresponds to the choice

$$u_k = \sqrt{\frac{\hbar}{2k}} e^{-ik\tau}\left(1 - \frac{i}{k\tau}\right), \tag{78}$$

where the coefficient has been determined by the normalization condition (76). This choice, in the jargon of quantum field theory on curved spacetimes, corresponds to the *Bunch–Davies vacuum*.

Finally, one can compute the *correlation function* for the scalar field ϕ in the vacuum state defined above. When Fourier transformed, the correlation function defines the *power spectrum* $\mathcal{P}_\phi(k)$:

$$\langle 0|\hat{\phi}(\mathbf{x}_1)\hat{\phi}(\mathbf{x}_2)|0\rangle = \int d^3k\ e^{i\mathbf{k}.(\mathbf{x}_1-\mathbf{x}_2)}\frac{\mathcal{P}_\phi(k)}{4\pi k^3}. \tag{79}$$

Note that the homogeneity and isotropy of the quantum field are used implicitly in the definition of the power spectrum, which is "diagonal" in Fourier space (homogeneity) and depends only on the norm of $\mathbf{k}$ (isotropy). In our case, we find

$$2\pi^2 k^{-3}\mathcal{P}_\phi = \frac{|u_k|^2}{a^2}, \tag{80}$$

which gives in the limit when the wavelength is *larger than the Hubble radius*, i.e. $k|\tau| \ll 1$,

$$\mathcal{P}_\phi(k) \simeq \hbar\left(\frac{H}{2\pi}\right)^2 \qquad (k \ll aH) \tag{81}$$

Note that, in the opposite limit, i.e. for wavelengths smaller than the Hubble radius ($k|\tau| \gg 1$), one recovers the usual result for fluctuations in Minkowski vacuum, $\mathcal{P}_\phi(k) = \hbar(k/2\pi a)^2$.

We have used a quantum description of the scalar field. But the cosmological perturbations are usually described by *classical random fields*. Roughly speaking, the transition between the quantum and classical (although stochastic) descriptions makes sense when the perturbations exit the Hubble radius. Indeed each of the terms in the Wronskian (76) is roughly of the order $\hbar/2(k\tau)^3$ in the super-Hubble limit and the non-commutativity can then be neglected. In this sense, one can see the exit outside the Hubble radius as a quantum-classical transition, although much refinement would be needed to make this statement more precise.

4.2 Quantum Fluctuations with Metric Perturbations

Let us now move to the more realistic case of a perturbed inflaton field living in a perturbed cosmological geometry. The situation is more complicated than in the previous problem, because Einstein's equations imply that scalar field fluctuations must necessarily coexist with *metric fluctuations*. A correct treatment, either classical or quantum, must therefore involve both the scalar field perturbations and metric perturbations. We thus need to resort to the theory of relativistic cosmological perturbations, which we briefly present below (more details can be found in, e.g., [28–30, 7, 31]).

4.2.1 Linear Perturbations of the Metric

The most general linear perturbation about the homogeneous metric can be expressed as

$$ds^2 = a^2\left\{-(1+2A)d\tau^2 + 2B_i dx^i d\tau + \left(\delta_{ij}+h_{ij}\right)dx^i dx^j\right\}, \tag{82}$$

where we have assumed, for simplicity, a spatially flat metric.[2]

We have introduced a time plus space decomposition of the perturbations. The indices i, j stand for *spatial* indices and the perturbed quantities defined in (82) can be seen as three-dimensional tensors, for which the indices can be lowered (or raised) by the spatial metric δ_{ij} (or its inverse).

It is very convenient to separate the perturbations into three categories, the so-called 'scalar', 'vector' and 'tensor' modes. For example, a spatial vector field B^i can be decomposed uniquely into a longitudinal part and a transverse part,

$$B_i = \partial_i B + \bar{B}_i, \quad \partial_i \bar{B}^i = 0, \tag{83}$$

where the longitudinal part is curl-free and can thus be expressed as a gradient, and the transverse part is divergenceless. One thus gets one "scalar" mode, B , and two "vector" modes $\bar{B}^i$ (the index i takes three values but the divergenceless condition implies that only two components are independent).

A similar procedure applies to the symmetric tensor h_{ij}, which can be decomposed as

$$h_{ij} = 2C\delta_{ij} + 2\partial_i\partial_j E + 2\partial_{(i}E_{j)} + \overline{E}_{ij}, \tag{84}$$

with $\overline{E}^{ij}$ transverse and traceless (TT), i.e. $\partial_i \overline{E}^{ij} = 0$ (transverse) and $\overline{E}^{ij}\delta_{ij} = 0$ (traceless), and E_i transverse. The parentheses around the indices denote symmetrization, namely $2\partial_{(i}E_{j)} = \partial_i E_j + \partial_j E_i$. We have thus defined two scalar modes: C and E, two vector modes, E_i, and two tensor modes, $\bar{E}_{ij}$.

4.2.2 Coordinate Transformations

The metric perturbations, introduced in (82), are modified in a coordinate transformation of the form

$$x^\alpha \to x^\alpha + \xi^\alpha, \qquad \xi^\alpha = (\xi^0, \xi^i). \tag{85}$$

It can be shown that the change of the metric components can be expressed as

$$\delta g_{\mu\nu} \to \delta g_{\mu\nu} - 2\nabla_{(\mu}\xi_{\nu)}, \tag{86}$$

where ∇ is the four-dimensional covariant derivative, and one considers the variation, due the coordinate transformation, at the *same* old and new coordinates (and thus at different physical points).

[2] This is moreover justified that the metric in the early Universe was closer to a spatially flat metric that our present metric which is indistinguishable from a flat geometry, according to observations.

The above variation can be decomposed into individual variations for the various components of the metric defined earlier. One finds

$$A \to A - \xi^{0'} - \mathcal{H}\xi^0 \tag{87}$$

$$B_i \to B_i + \partial_i \xi^0 - \xi_i' \tag{88}$$

$$h_{ij} \to h_{ij} - 2\left(\partial_{(i}\xi_{j)} - \mathcal{H}\xi^0\delta_{ij}\right), \tag{89}$$

where $\mathcal{H} \equiv a'/a$. The effect of a coordinate transformation can also be decomposed along the scalar, vector and tensor sectors introduced earlier. The generator ξ^α of the coordinate transformation can be written as

$$\xi^\alpha = (\xi^0, \partial^i \xi + \overline{\xi}^i), \tag{90}$$

with $\overline{\xi}^i$ transverse. This shows explicitly that ξ^α contains two scalar components, ξ^0 and ξ, and two vector components, $\overline{\xi}^i$. The transformations (88) and (89) are then decomposed into

$$\begin{aligned} B &\to B + \xi^0 - \xi', \\ C &\to C - \mathcal{H}\xi^0, \\ E &\to E - \xi, \\ \overline{B}^i &\to \overline{B}^i - \overline{\xi}^{i'}, \\ E^i &\to E^i - \overline{\xi}^i. \end{aligned} \tag{91}$$

The tensor perturbations remain unchanged since ξ^α does not contain any tensor component.

To summarize, the whole system scalar field plus gravitation is described by 11 perturbations. They can be decomposed into five scalar quantities: A, B, C and E from the metric and $\delta\phi$; four vector quantities $\bar{B}^i$ and $\bar{E}^i$; two tensor quantities, the two polarizations of E_{ij}^{TT}. However, these quantities are physically redundant since the same physical situation can be described by different sets of values of these perturbations, provided they are related by the coordinate transformations described above.

One would thus like to identify the true degrees of freedom, i.e. the physically independent quantities characterizing the system. Since spacetime has four coordinates, the coordinate transformations represent four gauge transformations, two of scalar type and two of vector types. Each gauge transformation reduces the number of degrees of freedom by two (the gauge "strikes" twice): one corresponds to the constraint associated with the gauge transformation, the other one because within the constrained hypersurface, one can fix the gauge. The situation for the scalar, vector and tensor sectors, respectively, is summarized in Table 1. The true degrees of freedom are thus the two polarizations of the gravitational waves and one scalar

Table 1 Counting of the degrees of freedom in the scalar, vector and tensor sectors

	Metric	Scalar field	Gauge choice	Constraints	True d.o.f.
S	4	1	−2	−2	1
V	4	0	−2	−2	0
T	2	0	0	0	2

degree of freedom. If matter was composed of N scalar fields, one would get N scalar degrees of freedom in addition to the two tensor modes.

In a coordinate transformation, the scalar field perturbation is also modified, according to

$$\delta\phi \to \delta\phi - \phi'\xi^0 \,. \tag{92}$$

In single-field inflation, there are thus two natural choices of gauge to describe the scalar perturbation. The first is to work with hypersurfaces that are flat, i.e. $C = 0$, in which case we will denote the scalar field perturbation by Q, i.e.

$$Q = \delta\phi_{C=0}. \tag{93}$$

The other choice is to work with hypersurfaces where the scalar field is uniform, i.e. $\delta\phi = 0$, in which case the scalar degree of freedom is embodied by the metric perturbation $C_{\delta\phi=0}$. In other words, the true scalar degree of freedom can be represented either as a pure matter perturbation or a pure metric perturbation. In the general case, we have [26, 32]

$$Q = \delta\phi - \frac{\phi'}{\mathcal{H}}C \,, \tag{94}$$

which is a gauge-invariant combination.

4.3 Quantizing the Scalar Degree of Freedom

In order to quantize the true scalar degree of freedom, one needs an action that governs its dynamics. Let us first note that the *linearized* equations of motion for the coupled system {gravity + scalar field} are obtained from the expansion of the full action at *second order* in the perturbations. Indeed the equations for the linear perturbations correspond to the Euler–Lagrange equations derived from a quadratic Lagrangian. In our case, the difficulty is that there are several scalar perturbations that are not independent. In order to quantize this coupled system, one can work directly with the second-order Lagrangian [30] or resort to a Hamiltonian approach [33, 34].

The modern approach, introduced by Maldacena [35] to study perturbations beyond linear order, is based on the Arnowitt–Deser–Misner (ADM) formalism [36],

which in some sense can be seen as a compromise between the purely Lagrangian and Hamiltonian approaches. In the ADM formalism, the metric is written in the form

$$ds^2 = -N^2 dt^2 + h_{ij}(dx^i + N^i dt)(dx^j + N^j dt), \tag{95}$$

where N is the lapse function and N^i the shift vector. The full action for the scalar field and for gravity

$$S_\phi = \int d^4x\sqrt{-g}\left(-\frac{1}{2}\partial^\mu\phi\partial_\mu\phi - V(\phi)\right) + \frac{M_P^2}{2}\int d^4x\sqrt{-g}R \tag{96}$$

becomes, after substitution of (95),

$$S = \int dt d^3x\sqrt{h}\left\{N\left[\frac{\mathcal{V}^2}{2N^2} - \frac{1}{2}h^{ij}\partial_i\phi\partial_j\phi - V(\phi)\right] + \frac{M_P^2}{2}\left[{}^{(3)}R + \frac{1}{N}(E_{ij}E^{ij} - E^2)\right]\right\}, \tag{97}$$

where ${}^{(3)}R$ is the scalar curvature of the spatial metric h_{ij} and h its determinant,

$$\mathcal{V} \equiv \dot{\phi} - N^j\partial_j\phi \tag{98}$$

and the symmetric tensor E_{ij}, defined by

$$E_{ij} \equiv \frac{1}{2}\dot{h}_{ij} - N_{(i|j)}\,, \tag{99}$$

(the symbol | denotes the spatial covariant derivative associated with the spatial metric h_{ij}) is proportional to the extrinsic curvature of the spatial slices.

The variation of the action with respect to N yields the energy constraint,

$$\frac{\mathcal{V}^2}{2N^2} + \frac{1}{2}h^{ij}\partial_i\phi\partial_j\phi + V(\phi) + \frac{M_P^2}{2N^2}(E_{ij}E^{ij} - E^2) - \frac{M_P^2}{2}(3)R = 0\,, \tag{100}$$

while the variation of the action with respect to the shift N_i gives the momentum constraint,

$$M_P^2\left(\frac{1}{N}(E_i^j - E\delta_i^j)\right)_{|j} = \frac{\mathcal{V}}{N}\partial_i\phi. \tag{101}$$

In order to study the linear perturbations about the FLRW background, we now restrict ourselves to the flat gauge, which corresponds to the choice

$$h_{ij} = a^2(t)\delta_{ij}. \tag{102}$$

The scalar fields on the corresponding flat hypersurfaces can be decomposed into

$$\phi = \bar{\phi} + Q, \tag{103}$$

where $\bar{\phi}$ is the spatially homogeneous background value of the scalar fields and Q represents its perturbation (in the flat case). In the following, we will often omit the bar and simply write the homogeneous value as ϕ, unless this could generate ambiguities.

We can also write the (scalarly) perturbed lapse and shift as

$$N = 1 + \alpha, \qquad N_i = \beta_{,i}\,, \tag{104}$$

where the linear perturbations α and β are determined in terms of the scalar field perturbations Q by solving the linearized constraints. At first order, the momentum constraint implies

$$\alpha = \frac{\dot{\phi}}{2M_P^2 H}\, Q\,, \tag{105}$$

while the energy constraint gives $\partial^2\beta$ in terms of Q and $\dot{Q}$.

4.4 Second-Order Action

We now expand, up to quadratic order, the action in terms of the linear perturbations. This action can be written solely in terms of the physical degree of freedom Q by substituting the expression (105) for α, since it turns out that β disappears from the second-order action, after an integration by parts. The second-order action can be written in the rather simple form

$$S_{(2)} = \frac{1}{2}\int dt\, \mathrm{d}^3x\, a^3 \left[\dot{Q}^2 - \frac{1}{a^2}\partial_i Q \partial^i Q - \mathcal{M}^2 Q^2\right], \tag{106}$$

with the effective (squared) mass

$$\mathcal{M}^2 = V'' - \frac{1}{a^3}\frac{d}{dt}\left(\frac{a^3}{H}\dot{\phi}^2\right). \tag{107}$$

As we did earlier, it is convenient to resort to conformal time τ and to introduce the canonical degree of freedom

$$v = a\, Q, \tag{108}$$

which leads to the action

$$S_v = \frac{1}{2}\int d\tau\ d^3x \left[v'^2 + \partial_i v \partial^i v + \frac{z''}{z} v^2\right] \tag{109}$$

with

$$z = a\,\frac{\phi'}{\mathcal{H}}. \tag{110}$$

This action is thus analogous to that of a scalar field in Minkowski spacetime with a time-dependent mass. One is thus back in a situation similar to the previous subsection, with the notable difference that the effective time-dependent mass is now z''/z, instead of a''/a.

The quantity we will be eventually interested in is the comoving curvature perturbation $\mathcal{R}$, which is related to the canonical variable v by the relation

$$v = z\,\mathcal{R}. \tag{111}$$

Since, by analogy with (80), the power spectrum for v is given by

$$2\pi^2 k^{-3}\mathcal{P}_v(k) = |v_k|^2, \tag{112}$$

the corresponding power spectrum for $\mathcal{R}$ is found to be

$$2\pi^2 k^{-3}\mathcal{P}_{\mathcal{R}}(k) = \frac{|v_k|^2}{z^2}. \tag{113}$$

In the case of an inflationary phase in the *slow-roll* approximation, the evolution of ϕ and of H is much slower than that of the scale factor a. Consequently, one gets approximately

$$\frac{z''}{z} \simeq \frac{a''}{a}, \qquad \text{(slow-roll)} \tag{114}$$

and all results of the previous section obtained for u apply directly to our variable v in the slow-roll approximation. This implies that the properly normalized function corresponding to the Bunch–Davies vacuum is approximately given by

$$v_k \simeq \sqrt{\frac{\hbar}{2k}}e^{-ik\tau}\left(1 - \frac{i}{k\tau}\right). \tag{115}$$

In the super-Hubble limit $k|\tau| \ll 1$, the function v_k behaves like

$$v_k \simeq -\sqrt{\frac{\hbar}{2k}}\,\frac{i}{k\tau} \simeq i\sqrt{\frac{\hbar}{2k}}\,\frac{aH}{k}, \tag{116}$$

where we have used $a \simeq -1/(H\tau)$.

Consequently, combining (113), (110) and (115) and reintroducing the cosmic time gives the power spectrum for $\mathcal{R}$, on scales larger than the Hubble radius,

$$\mathcal{P}_{\mathcal{R}} \simeq \frac{\hbar}{4\pi^2}\left(\frac{H^4}{\dot{\phi}^2}\right)_{k=aH} = \frac{\hbar}{2M_P^2\varepsilon_*}\left(\frac{H_*}{2\pi}\right)^2 \tag{117}$$

where we have used $\varepsilon \equiv -\dot{H}/H^2$ in the second equality, and the subscript $*$ means that the quantity is evaluated at Hubble crossing ($k = aH$). This is the main result for the spectrum of scalar cosmological perturbations generated from vacuum fluctuations during a slow-roll inflation phase.

4.5 Gravitational Waves

We have focused so far our attention on scalar perturbations, which are the most important in cosmology. Tensor perturbations, or primordial gravitational waves, if ever detected in the future, would be a remarkable probe of the early Universe. In the inflationary scenario, like scalar perturbations, primordial gravitational waves are generated from vacuum quantum fluctuations [37]. Let us now explain briefly this mechanism.

The action expanded at second order in the perturbations contains a tensor part given by

$$S_g^{(2)} = \frac{M_P^2}{8}\int d\tau\, d^3x\, a^2 \eta^{\mu\nu}\partial_\mu \bar{E}_j^i \partial_\nu \bar{E}_i^j, \tag{118}$$

where $\eta^{\mu\nu}$ denotes the Minkowski metric. Apart from the tensorial nature of E_j^i, this action is quite similar to that of a scalar field in a FLRW Universe (65), up to a renormalization factor $M_P/2$. The decomposition

$$a\bar{E}_j^i = \sum_{\lambda=+,\times}\int \frac{d^3k}{(2\pi)^{3/2}} v_{k,\lambda}(\tau)\varepsilon_j^i(\mathbf{k};\lambda)e^{i\mathbf{k}.\mathbf{x}}, \tag{119}$$

where the $\varepsilon_j^i(\mathbf{k};\lambda)$ are the polarization tensors, shows that the gravitational waves are essentially equivalent to two massless scalar fields (for each polarization) $\phi_\lambda = M_P\bar{E}_\lambda/2$.

The total power spectrum is thus immediately deduced from (80):

$$\mathcal{P}_T = 2 \times \frac{4}{M_P^2} \times \hbar\left(\frac{H}{2\pi}\right)^2, \tag{120}$$

where the first factor comes from the two polarizations, the second from the renormalization with respect to a canonical scalar field, the last term being the power spectrum for a scalar field derived earlier. In summary, the tensor power spectrum is

$$\mathcal{P}_T = \frac{8\hbar}{M_P^2}\left(\frac{H_*}{2\pi}\right)^2, \tag{121}$$

where the subscript $*$ recalls that the Hubble parameter, which can be slowly evolving during inflation, must be evaluated when the relevant scale exited the Hubble radius during inflation.

A measurement of the tensor amplitude (121) gives direct access, in this context, to the energy scale H_* during inflation, in contrast with the scalar amplitude (117) which depends on the slow-roll parameter, ε_*, as well. The tensor to scalar ratio,

$$r \equiv \frac{\mathcal{P}_T}{\mathcal{P}_\mathcal{R}} = 16\,\varepsilon_*\,, \tag{122}$$

is proportional to the slow-roll parameter.

5 From Inflation to the Standard Era

Once the perturbations have been computed during inflation, it is necessary to relate them to the perturbations in the standard phase of cosmology, where they will be used as "initial conditions". A priori, one could think that it is necessary to follow the details of how the inflaton is converted into ordinary particles in order to establish this relation. In fact, all these details turn out to be irrelevant, at least in the case of single inflation, because all the scales of cosmological interest are larger than the Hubble radius at the end of inflation and there exists a conservation law on super-Hubble scales, as we will see in this section.

5.1 Covariant Approach

Instead of the traditional metric-based approach, we use here a more geometrical approach to cosmological perturbations [38], which will enable us to recover easily and intuitively the main useful results, not only for linear perturbations but also for non-linear perturbations.

Let us thus consider a spacetime with metric g_{ab} and some perfect fluid characterized by its energy density ρ, its pressure P and its four-velocity u^a. The corresponding energy–momentum tensor is given by

$$T_{ab} = \rho u_a u_b + P(g_{ab} + u_a u_b). \tag{123}$$

Let us also introduce the expansion along the fluid worldlines,

$$\Theta = \nabla_a u^a, \tag{124}$$

and the integrated expansion

$$\alpha = \frac{1}{3}\int d\tau\, \Theta\ , \tag{125}$$

where τ is the proper time defined along the fluid worldlines. In a FLRW spacetime, one would find $\Theta = 3H$. Therefore, in the general case, one can interpret $\Theta/3$ as a local Hubble parameter and $S = \exp(\alpha)$ as a local scale factor, while α represents the local number of e-folds.

As shown in [39, 40], the conservation law for the energy–momentum tensor,

$$\nabla_a T^a_{\ b} = 0, \tag{126}$$

implies that the covector

$$\zeta_a \equiv \nabla_a \alpha - \frac{\dot{\alpha}}{\dot{\rho}}\nabla_a \rho \tag{127}$$

satisfies the relation

$$\dot{\zeta}_a \equiv \mathcal{L}_u \zeta_a = -\frac{\Theta}{3(\rho + p)}\left(\nabla_a p - \frac{\dot{p}}{\dot{\rho}}\nabla_a \rho\right)\ , \tag{128}$$

where a dot denotes a Lie derivative along u^a, which is equivalent to an ordinary derivative for *scalar* quantities (e.g. $\dot{\rho} \equiv u^a \nabla_a \rho$). This result is valid for any spacetime geometry and does not depend on Einstein's equations. In the cosmological context, α can be interpreted as a non-linear generalization, according to an observer following the fluid, of the number of e-folds of the scale factor.

The covector ζ_a can be defined for the global cosmological fluid or for any of the individual cosmological fluids (the case of interacting fluids is discussed in [41]). Using the non-linear conservation equation

$$\dot{\rho} = -3\dot{\alpha}(\rho + P)\ , \tag{129}$$

which follows from $u^b \nabla_a T^a_{\ b} = 0$, one can re express ζ_a in the form

$$\zeta_a = \nabla_a \alpha + \frac{\nabla_a \rho}{3(\rho + P)}\ . \tag{130}$$

If $w \equiv P/\rho$ is constant, the above covector is a total gradient and can be written as

$$\zeta_a = \nabla_a\left[\alpha + \frac{1}{3(1+w)}\ln\rho\right]\ . \tag{131}$$

On scales larger than the Hubble radius, the above definitions are equivalent to the non-linear curvature perturbation on uniform density hypersurfaces as defined in [42] (see also [43])

$$\zeta = \delta N - \int_{\bar{\rho}}^{\rho} H \frac{d\tilde{\rho}}{\dot{\tilde{\rho}}} = \delta N + \frac{1}{3}\int_{\bar{\rho}}^{\rho} \frac{d\tilde{\rho}}{(1+w)\tilde{\rho}}\,, \tag{132}$$

where $N = \alpha$. The above equation is simply the integrated version of (127) or of (130).

5.2 Linear Conserved Quantities

Let us now introduce a coordinate system, in which the metric (with scalar perturbations) reads

$$ds^2 = a^2\left\{-(1+2A)d\tau^2 + 2\partial_i B\, dx^i d\tau + \left[(1+2C)\delta_{ij} + 2\partial_i\partial_j E\right] dx^i dx^j\right\}. \tag{133}$$

One can decompose the fluid four-velocity as

$$u^\mu = \bar{u}^\mu + \delta u^\mu, \quad \delta u^\mu = \left\{-A/a, v^i/a\right\}, \quad v_i = \partial_i v + \bar{v}_i\,, \tag{134}$$

where $\bar{v}_i$ is transverse.

At linear order, the spatial components of ζ_a are simply

$$\zeta_i^{(1)} = \partial_i \zeta^{(1)}, \qquad \zeta^{(1)} \equiv \delta\alpha - \frac{\bar{\alpha}'}{\bar{\rho}'}\delta\rho. \tag{135}$$

Linearizing (128) implies that the curvature perturbation on *uniform energy density hypersurfaces*, defined by

$$\zeta = C - \mathcal{H}\frac{\delta\rho}{\rho'} = C + \frac{\delta\rho}{3(\rho+P)} \tag{136}$$

and originally introduced in [44], obeys the evolution equation (see also [45])

$$\zeta' = -\frac{\mathcal{H}}{\rho+P}\delta P_{\rm nad} - \frac{1}{3}\nabla^2(E'+v), \tag{137}$$

where $\delta P_{\rm nad}$ is the non-adiabatic part of the pressure perturbation, defined by

$$\delta P_{\rm nad} = \delta P - c_s^2\delta\rho. \tag{138}$$

Note that $\zeta^{(1)}$ differs from ζ but they coincide when the spatial gradients can be neglected, for instance, on large scales. The expression (137) shows that ζ is conserved *on super-Hubble scales* in the case of *adiabatic* perturbations.

Another convenient quantity, which is sometimes used in the literature instead of ζ, is the *curvature perturbation on comoving hypersurfaces*, which can be written in any gauge as

$$\mathcal{R} = -C - \frac{\mathcal{H}}{\rho + P}\delta q, \qquad \partial_i \delta q \equiv \delta_{(S)} T_i^0, \tag{139}$$

where the subscript (S) denotes the perturbations of scalar type. For a perfect fluid, $\delta q = (\rho + P)(v + B)$, where v has been defined in (134).

One can relate the two quantities ζ and $\mathcal{R}$ by using the energy and momentum constraints, which were derived earlier in the ADM formalism. Linearizing (100) and (101) yields, respectively,

$$3\mathcal{H}^2\delta N + a\mathcal{H}\partial^2\beta = -\frac{a^3}{2M_P^2}\delta\rho, \tag{140}$$

$$\mathcal{H}\delta N = -\frac{a^3}{2M_P^2}\delta q\,. \tag{141}$$

Combining these two equations yields the relativistic analog of the Poisson equation, namely

$$\partial^2\Psi = \frac{a^2}{2M_P^2}(\delta\rho - 3\mathcal{H}\delta q) \equiv \frac{a^2}{2M_P^2}\delta\rho_c\,, \tag{142}$$

where we have replaced β by the Bardeen potential $\Psi \equiv -C - \mathcal{H}(B - E') = -\mathcal{H}\beta$ and introduced the comoving energy density $\delta\rho_c \equiv \delta\rho - 3\mathcal{H}\delta q$. Since

$$\zeta = -\mathcal{R} + \frac{\delta\rho_c}{\rho + P} = -\mathcal{R} - \frac{2\rho}{3(\rho + P)}\left(\frac{k}{aH}\right)^2\Psi, \tag{143}$$

one finds that ζ and $\mathcal{R}$ coincide in the super-Hubble limit $k \ll aH$.

5.3 "Initial" Conditions for Standard Cosmology

In standard cosmology, the "initial" conditions for the perturbations are usually defined in the radiation-dominated era around the time of nucleosynthesis, when the main cosmological components are restricted to the usual photons, baryons, neutrinos and cold dark matter (CDM) particles. The scales that are cosmologically relevant today correspond to lengthscales much larger than the Hubble radius at that time.

Before inflation, the "initial" conditions were put "by hand", with the restriction that their late-time consequences should be compatible with observations. Inflation now provides a precise prescription to determine these "initial conditions".[3]

[3] Although one must be aware that present cosmological scales can correspond to scales smaller than the Planck scale during inflation, suggesting the possibility of trans-Planckian effects (see, e.g. [46]).

Since several species are present, one needs to specify the density perturbation of each species. A simplification arises in the case of *single-field* inflation, since exactly the same cosmological history, i.e. inflation followed by the decay of the inflation into the usual species, occurs in all parts of our Universe, up to a small time shift depending on the perturbation of the inflaton in each region. As a consequence, even if the number densities of the various species vary from point to point, their ratio should be fixed, i.e.

$$\delta\left(n_A/n_B\right) = 0\,, \tag{144}$$

for any pair of species denoted A and B (see, e.g. [47] for a more detailed discussion). This is not necessarily true in *multi-field inflation*, as the perturbations in the radiation era may depend on *different* combinations of the scalar field perturbations.

The variation in the relative particle number densities between two species can be quantified by the quantity

$$S_{A,B} \equiv \frac{\delta n_A}{n_A} - \frac{\delta n_B}{n_B}\,, \tag{145}$$

which is usually called the *entropy* perturbation between A and B. When the equation of state for a given species is such that $w \equiv P/\rho = \text{const}$, one can reexpress the entropy perturbation in terms of the density contrast, in the form

$$S_{A,B} \equiv \frac{\delta_A}{1+w_A} - \frac{\delta_B}{1+w_B}. \tag{146}$$

It is convenient to choose a species of reference, for instance, the photons, and to define the entropy perturbations of the other species relative to it. The quantities

$$S_b \equiv \delta_b - \tfrac{3}{4}\delta_\gamma, \tag{147}$$

$$S_c \equiv \delta_c - \tfrac{3}{4}\delta_\gamma, \tag{148}$$

$$S_\nu \equiv \tfrac{3}{4}\delta_\nu - \tfrac{3}{4}\delta_\gamma, \tag{149}$$

thus define, respectively, the baryon, CDM and neutrino entropy perturbations.

For single-field inflation, all these entropy perturbations vanish, $S_b = S_c = S_\nu = 0$, and the primordial perturbations are said to be *adiabatic*. An adiabatic primordial perturbation is thus characterized by

$$\frac{1}{4}\delta_\gamma = \frac{1}{4}\delta_\nu = \frac{1}{3}\delta_b = \frac{1}{3}\delta_c\,. \tag{150}$$

Only one density constrast needs to be specified. However, since it is a gauge-dependent quantity, it is better to use the gauge-invariant quantity ζ, i.e. the uniform density curvature perturbation, which is also equivalent to $-\mathcal{R}$, since we are on super-Hubble scales here.

Note that the *adiabatic mode*, which is directly related to the curvature perturbation, is also called *curvature* mode. By contrast, the entropy perturbations can be non-zero even if the curvature is zero, and the corresponding modes are called *isocurvature* modes.

5.4 Inflation and Cosmological Data

Let us now discuss the confrontation of single-field inflation models with the current cosmological data. The main idea is that one can predict precisely the statistics of the CMB perturbations, once the amplitude of the primordial perturbation as a function of scale, $\mathcal{R}(k)$, is given, provided some choice for the cosmological parameters Ω_i. In other words, the measurements of the CMB, together with other cosmological data, allow us to constrain both the cosmological parameters, which are numbers, and the primordial spectrum, which is a function (see, e.g. [48, 49] for details on the CMB physics). From the present data, one finds that the primordial spectrum is nearly (although not quite) scale invariant , with an amplitude

$$\mathcal{P}_{\mathcal{R}}^{1/2} \simeq 5 \times 10^{-5}. \tag{151}$$

In order to derive some constraints on the inflation models, it is useful to reexpress the scalar and tensor power spectra, respectively, given in (117) and (121), in terms of the scalar field potential. This can be done by using the slow-roll equations (29) and (30). One finds for the scalar spectrum

$$\mathcal{P}_{\mathcal{R}} = \frac{1}{12\pi^2}\left(\frac{V^3}{M_P^6 V'^2}\right)_{k=aH} \tag{152}$$

with subscript meaning that the term on the right-hand side must be evaluated at *Hubble radius exit* for the scale of interest. The scalar spectrum can also be written in terms of the first slow-roll parameter defined in (32), in which case it reads

$$\mathcal{P}_{\mathcal{R}} = \frac{1}{24\pi^2}\left(\frac{V}{M_P^4 \varepsilon_V}\right)_{k=aH}. \tag{153}$$

If ε_V is order 1, as in chaotic models, the observed amplitude (152) implies that the typical energy scale during inflation is

$$V^{1/4} \sim 10^{-3} M_P \sim 10^{15}\text{GeV}. \tag{154}$$

The tensor power spectrum, in terms of the scalar field potential, is given by

$$\mathcal{P}_T = \frac{2}{3\pi^2}\left(\frac{V}{M_P^4}\right)_{k=aH}. \tag{155}$$

The scalar and tensor spectra are almost scale invariant but not quite since the scalar field evolves slowly during the inflationary phase. In order to evaluate quantitatively this variation, it is convenient to introduce a scalar *spectral index* as well as a tensor one, defined, respectively, by

$$n_S(k) - 1 = \frac{d\ln\mathcal{P}_{\mathcal{R}}(k)}{d\ln k}, \quad n_T(k) = \frac{d\ln\mathcal{P}_T(k)}{d\ln k}. \tag{156}$$

One can express the spectral indices in terms of the slow-roll parameters. For this purpose, let us note that, in the slow-roll approximation, $d\ln k = d\ln(aH) \simeq d\ln a$, so that

$$\frac{d\phi}{d\ln a} = \frac{\dot{\phi}}{H} \simeq -\frac{V'}{3H^2} \simeq -M_P^2\frac{V'}{V}, \tag{157}$$

where the slow-roll equations (29) and (30) have been used. Therefore, one gets

$$n_s(k) - 1 = 2\eta_V - 6\varepsilon_V, \tag{158}$$

where ε_V and η_V are the two slow-roll parameters defined in (32) and (35). Similarly, one finds for the tensor spectral index

$$n_T(k) = -2\varepsilon_V. \tag{159}$$

Comparing with (122), this yields the relation

$$r = -8n_T, \tag{160}$$

the so-called *consistency relation* which relates purely *observable* quantities. This means that if one was able to observe the primordial gravitational waves and measure the amplitude and spectral index of their spectrum, a rather formidable task, then one would be able to test directly the paradigm of single-field slow-roll inflation.

Let us now discuss the particular models which we have already considered, and let us establish the predictions of these models in a (n_s,r) plane, where they can be directly compared with the observational constraints. For the power law potentials (41), one finds, using (42),

$$n_s - 1 = -6\varepsilon + 2\eta = -2\frac{p+2}{p}\varepsilon \tag{161}$$

and

$$r = 16\varepsilon = \frac{8p}{p+2}(1 - n_s). \tag{162}$$

Moreover,

$$\varepsilon = \frac{p}{4N}, \tag{163}$$

where N is the number of e-folds before the end of inflation when the scales of cosmological interest crossed out the Hubble radius. Therefore, the observational prediction for a model with a power law potential lie on a line in the (n_s,r) plane, the precise point depending on the number of e-folds when the perturbations were generated.

For an exponential potential (49), one finds, using (52),

$$n_s - 1 = -\frac{2}{q} \tag{164}$$

and

$$r = \frac{16}{q}. \tag{165}$$

The prediction in the (n_s,r) plane thus depends only on the parameter in the exponential of the potential, but not on the number of e-folds as in the previous case.

For potentials (53) like in hybrid inflation, one finds

$$\eta = \frac{1 + \tilde{\phi}^2}{\tilde{\phi}^2}\varepsilon \tag{166}$$

and

$$r = 8\frac{\tilde{\phi}^2}{2\tilde{\phi}^2 - 1}(1 - n_s). \tag{167}$$

One can proceed in a similar way for any model of inflation and thus be able to confront it with observational data.

Before concluding this section, it is worth noticing that a significant amount of gravitational waves, and thus a detectable r, requires a variation of the inflaton of the order of the Planck mass during inflation [50].

6 More General Inflationary Scenarios

So far, the simplest models of inflation are compatible with the data but it is instructive to study more refined models for at least two reasons. First, models inspired by high-energy physics are usually more complicated than the simplest

phenomenological inflationary models. Second, exploring more general models of inflation and identifying their specific observational features is a healthy procedure to prepare the interpretation of the future data.

At present, two types of extensions of the simplest scenarios have been mainly studied:

- models with non-standard kinetic terms;
- models with multiple scalar fields.

Of course, the two aspects can be combined and there exist scenarios involving several scalar fields with non-standard kinetic terms, as we will see later.

Among the scenarios involving several scalar fields, it is useful to distinguish three subclasses. The first, and oldest, category includes the models with *multiple inflatons*, i.e. models where several scalar fields play a dynamical role in the homogeneous cosmological evolution during inflation. In the second category, one finds the *curvaton* scenarios. These models assume the existence, in addition to the inflaton, of a scalar field, which is light during inflation but does not participate to inflation per se. Its energy density, which decreases less quickly than radiation, becomes significant only after inflation. Its decay produces a second reheating, and its fluctuations are then imprinted in the curvature perturbation.

The final subclass regroups are what we will name as the *modulaton* scenarios. Like in the curvaton models, one assumes the presence of a light scalar field, the modulaton, which is subdominant during inflation but acquires some fluctuations. The fluctuations of the modulaton are transferred to the curvature perturbation because the cosmological evolution is governed by some parameter that depends on the modulaton. This parameter can be, for instance, the value of the inflaton at the end of inflation, or the coupling of the inflaton to other particles during the reheating. Of course, one can envisage even more complicated scenarios which combine several of these mechanisms.

6.1 Generalized Lagrangians

We now consider multi-field models, which can be described by an action of the form

$$S = \int d^4x\sqrt{-g}\left[\frac{R}{16\pi G} + P(X^{IJ},\phi^K)\right], \tag{168}$$

where P is an arbitrary function of N scalar fields and of the kinetic term

$$X^{IJ} = -\frac{1}{2}\nabla_\mu\phi^I\nabla^\mu\phi^J. \tag{169}$$

The relations obtained earlier for the single-field model can then be generalized. The energy–momentum tensor, derived from (168), is of the form

$$T^{\mu\nu} = Pg^{\mu\nu} + P_{<IJ>}\partial^{\mu}\phi^{I}\partial^{\nu}\phi^{J}\,, \tag{170}$$

where $P_{<IJ>}$ denotes the partial derivative of P with respect to X^{IJ} (symmetrized with respect to the indices I and J). The equations of motion for the scalar fields, which can be seen as generalized Klein–Gordon equations, are obtained from the variation of the action with respect to ϕ^{I}. One finds

$$\nabla_{\mu}\left(P_{<IJ>}\nabla^{\mu}\phi^{J}\right) + P_{,I} = 0, \tag{171}$$

where $P_{,I}$ denotes the partial derivative of P with respect to ϕ^{I}.

In a homogeneous spacetime, $X^{IJ} = \dot{\phi}^{I}\dot{\phi}^{J}/2$, and the energy–momentum tensor reduces to that of a perfect fluid with energy density

$$\rho = 2P_{<IJ>}X^{IJ} - P\,, \tag{172}$$

and pressure P. The evolution of the scale factor $a(t)$ is governed by the Friedmann equations, which can be written in the form

$$H^2 = \frac{1}{3}\left(2P_{<IJ>}X^{IJ} - P\right)\,, \qquad \dot{H} = -X^{IJ}P_{<IJ>}\,. \tag{173}$$

The equations of motion for the scalar fields reduce to

$$\left(P_{<IJ>} + P_{<IL>,<JK>}\dot{\phi}^{L}\dot{\phi}^{K}\right)\ddot{\phi}^{J} + \left(3HP_{<IJ>} + P_{<IJ>,K}\dot{\phi}^{K}\right)\dot{\phi}^{J} - P_{,I} = 0\,, \tag{174}$$

where $P_{<IL>,<JK>}$ denotes the (symmetrized) second derivative of P with respect to X^{IL} and X^{JK}.

The expansion up to second order in the linear perturbations of the action (168) is useful to obtain the classical equations of motion for the perturbations and to calculate the spectra of the primordial perturbations generated during inflation, as we have seen earlier in the case of a single scalar field. Working for convenience with the scalar field perturbations Q^{I} defined in the spatially flat gauge, the second-order action can be written in the compact form [51]:

$$S_{(2)} = \frac{1}{2}\int dt\, d^3x\, a^3\left[\left(P_{<IJ>} + 2P_{<MJ>,<IK>}X^{MK}\right)\dot{Q}^{I}\dot{Q}^{J}\right.$$
$$\left. - P_{<IJ>}h^{ij}\partial_i Q^{I}\partial_j Q^{J} - \mathcal{M}_{KL}Q^{K}Q^{L} + 2\,\Omega_{KI}Q^{K}\dot{Q}^{I}\right], \tag{175}$$

where the mass matrix is

$$\mathcal{M}_{KL} = -P_{,KL} + 3X^{MN}P_{<NK>}P_{<ML>} + \frac{1}{H}P_{<NL>}\dot{\phi}^N\left[2P_{<IJ>,K}X^{IJ} - P_{,K}\right] - \frac{1}{H^2}X^{MN}P_{<NK>}P_{<ML>}\left[X^{IJ}P_{<IJ>} + 2P_{<IJ>,<AB>}X^{IJ}X^{AB}\right] - \frac{1}{a^3}\frac{d}{dt}\left(\frac{a^3}{H}P_{<AK>}P_{<LJ>}X^{AJ}\right) \tag{176}$$

and the mixing matrix is

$$\Omega_{KI} = \dot{\phi}^J P_{<IJ>,K} - \frac{2}{H}P_{<LK>}P_{<MJ>,<NI>}X^{LN}X^{MJ}\,. \tag{177}$$

This formalism is very general and it is instructive to consider two particular cases, which have often been studied in the literature.

6.2 Simple Multi-inflaton Scenarios

The first category includes multi-field scenarios governed by a Lagrangian of the form

$$P = G_{IJ}X^{IJ} - V = -\frac{1}{2}G_{IJ}(\phi)\,\partial^\mu\phi^I\partial_\mu\phi^J - V(\phi)\,, \tag{178}$$

where the field metric G_{IJ} can be non-trivial (also studied in, e.g. [52–54]) It can then be shown that the second-order action can be rewritten in the form

$$S_{(2)} = \frac{1}{2}\int dt\,\mathrm{d}^3x\,a^3\left[G_{IJ}\mathcal{D}_tQ^I\mathcal{D}_tQ^J - \frac{1}{a^2}G_{IJ}\partial_iQ^I\partial^iQ^J - \tilde{M}_{IJ}Q^IQ^J\right], \tag{179}$$

with

$$\tilde{M}_{IJ} = \mathcal{D}_I\mathcal{D}_JV - R_{IKLJ}\dot{\phi}^K\dot{\phi}^L - \frac{1}{a^3}\mathcal{D}_t\left[\frac{a^3}{H}\dot{\phi}_I\dot{\phi}_J\right], \tag{180}$$

where $\mathcal{D}_I$ denotes the covariant derivative with respect to the field space metric G_{IJ} (so that $\mathcal{D}_I\mathcal{D}_JV = V_{,IJ} - \Gamma^K_{IJ}V_{,K}$ where Γ^K_{IJ} denote the Christoffel symbols of the metric G_{IJ}), R_{IJKL} is the corresponding Riemann tensor and $\mathcal{D}_tQ^I \equiv \dot{Q}^I + \Gamma^I_{JK}\dot{\phi}^IQ^K$.

It is now convenient, following the approach of [55], to introduce a particular direction in field space, which we will call the *instantaneous adiabatic* direction, defined by the unit vector tangent to the inflationary trajectory in field space,

$$e^I_\sigma = \frac{\dot{\phi}^I}{\sqrt{2X}} = \frac{\dot{\phi}^I}{\dot{\sigma}}\,, \tag{181}$$

where we have introduced the notation $X \equiv G_{IJ}X^{IJ}$ and $\dot{\sigma} \equiv \sqrt{2X}$.

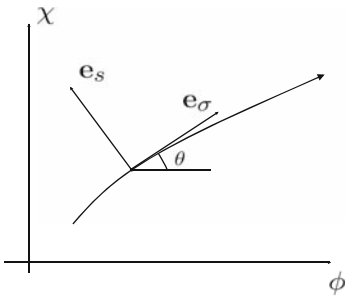

Fig. 3 Inflationary trajectory in a two-field model. The (instantaneous) adiabatic vector e_σ is tangent to the trajectory while the (instantaneous) entropic vector e_s is orthogonal to it

The directions orthogonal to e^I_σ are called *instantaneous entropic* and span an hyperplane in field space. For simplicity, let us now concentrate on two-field scenarios, where there is a single entropic degree of freedom (Fig. 3). Defining the entropy vector e^I_s as the unit vector orthogonal to the adiabatic vector e^I_σ, i.e.

$$G_{IJ}e^I_s e^J_s = 1, \qquad G_{IJ}e^I_s e^J_\sigma = 0, \tag{182}$$

one can then uniquely decompose the scalar field perturbations into (instantaneous) adiabatic and entropic modes,

$$Q^I = Q_\sigma e^I_\sigma + Q_s e^I_s \,. \tag{183}$$

One can derive the equations of motion for the quantities Q_σ and Q_s from the second-order action. One finds [54]

$$\ddot{Q}_\sigma + 3H\dot{Q}_\sigma + \left(\frac{k^2}{a^2} + \mu^2_\sigma\right) Q_\sigma = (\Xi Q_s)^\cdot - \left(\frac{\dot{H}}{H} + \frac{V_{,\sigma}}{\dot{\sigma}}\right) \Xi\, Q_s \,, \tag{184}$$

with

$$\Xi \equiv -\frac{2}{\dot{\sigma}} V_{,s} \,, \qquad \mu^2_\sigma \equiv -\frac{(\dot{\sigma}/H)^{\cdot\cdot}}{\dot{\sigma}/H} - \left(3H + \frac{(\dot{\sigma}/H)^\cdot}{\dot{\sigma}/H}\right) \frac{(\dot{\sigma}/H)^\cdot}{\dot{\sigma}/H}, \tag{185}$$

where $V_{,\sigma} \equiv e^I_\sigma V_{,I}$ and $V_{,s} \equiv e^I_s V_{,I}$. The equation of motion for the entropy part is given by

$$\ddot{Q}_s + 3H\dot{Q}_s + \left(\frac{k^2}{a^2} + \mu^2_s\right) Q_s = -\Xi \left[\dot{Q}_\sigma - H\left(\frac{\dot{\sigma}^2}{2H^2} + \frac{\ddot{\sigma}}{H\dot{\sigma}}\right) Q_\sigma\right], \tag{186}$$

with

$$\mu^2_s \equiv V_{ss} + \frac{1}{2}\dot{\sigma}^2 R - \frac{V^2_{,s}}{2X} \,, \tag{187}$$

where R is the trace of the Ricci tensor on field space, i.e. the scalar curvature.

The adiabatic perturbation is naturally related to the comoving curvature perturbation (139). Indeed, using the energy–momentum tensor (170), with the property $\rho + P = 2X$, which follows from (172), one finds that the comoving perturbation (139) is given by

$$\mathcal{R} = \frac{H}{2X}\dot{\phi}_I Q^I = \frac{H}{\sqrt{2X}} Q_\sigma \,. \tag{188}$$

Taking the time derivative of this expression and using the analog of (142),

$$-2\frac{k^2}{a^2}\Psi = \delta\rho_c = \sqrt{2X}\left[\dot{Q}_\sigma + \left(\frac{\dot{H}}{H} - \frac{\dot{X}}{2X}\right) Q_\sigma\right] + 2V_{,s}Q_s \tag{189}$$

one finds

$$\dot{\mathcal{R}} = \frac{H}{\dot{H}}\frac{k^2}{a^2}\Psi - 2\frac{H}{\dot{\sigma}^2}V_{,s}Q_s \,. \tag{190}$$

By noting that the right-hand side of (186) is proportional to $\dot{\mathcal{R}}$, one can rewrite the entropic equation of motion as

$$\ddot{Q}_s + 3H\dot{Q}_s + \left(\frac{k^2}{a^2} + \mu_s^2 + \Xi^2\right) Q_s = -\frac{\dot{\sigma}}{\dot{H}}\Xi\frac{k^2}{a^2}\Psi \,. \tag{191}$$

When the spatial gradients can be neglected on large scales, the above equation shows that the entropy perturbation Q_s evolves independently of the adiabatic mode. In the same limit, the adiabatic mode is governed by a *first-order* equation

$$\dot{\mathcal{R}} \approx \frac{H}{\dot{\sigma}}\,\Xi\, Q_s \quad \text{or} \quad \dot{Q}_\sigma + \left(\frac{\dot{H}}{H} - \frac{\ddot{\sigma}}{\dot{\sigma}}\right) Q_\sigma - \Xi\, Q_s \approx 0. \tag{192}$$

This implies that, in contrast with the entropy mode, the adiabatic mode is affected by the entropy on large scales, as soon as the *mixing parameter* $\Xi = -2V_{,s}/\dot{\sigma}$ is non zero. When the field metric is flat, $G_{IJ} = \delta_{IJ}$, one can introduce the rotation angle between the initial basis and the adiabatic/entropy basis. One thus finds that $\Xi = 2\dot{\theta}$. In the case a field metric of the form

$$G_{IJ}\, d\phi^I\, d\phi^J = d\phi^2 + e^{2b(\phi)} d\chi^2 \,, \tag{193}$$

investigated in [56, 57], the coupling is given by $\Xi = 2\dot{\theta} + b'\dot{\sigma}\sin\theta$, where the additional term simply comes from the non-trivial covariant derivative. Note that non-linear extensions of the adiabatic and entropic equations have been obtained in [58, 59] (see also [60, 61] for other works on non-linear perturbations in multi-field inflation).

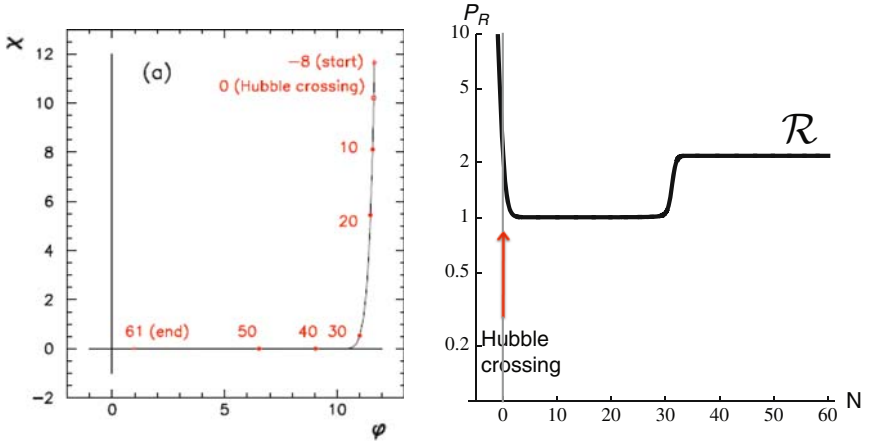

Fig. 4 In a double inflation model, with two different masses for the scalar fields, the inflationary trajectory is bent (*left*). This induces an evolution of the curvature perturbation, *after* Hubble crossing (*right*). Other examples can be found in [57]

The above results show that a generic feature of multi-inflaton scenarios is that the curvature perturbation is not frozen at horizon crossing, like in single-field inflation, but can, instead, evolve on large scales as a consequence of the *transfer* of entropy perturbations into adiabatic perturbations, as illustrated in Fig. 4. This was pointed out originally in [62] in the context of generalized gravity theories. It is thus crucial, when working with a model involving several scalar fields during inflation, to identify all the light directions in field space and to evolve the curvature perturbation until any transfer from entropy into adiabatic modes has completely ceased (the transfer can even occur long after inflation, as is the case in the curvaton scenario, which we will discuss later).

As we have just seen, the instantaneous entropy perturbations can affect the evolution of the curvature perturbation during inflation, on large scales, but they could also survive the end of inflation and the reheating phase and therefore, cause the existence of "initial" isocurvature perturbations, for instance, between the CDM and photon fluids, in the radiation era. Moreover, these isocurvature perturbations could be correlated with the "initial" adiabatic perturbations [63], since part of the adiabatic perturbation can originate from an (instantaneous) entropy perturbation during inflation. We will discuss later the observational constraints on this possibility.

6.3 K-Inflation

Let us now consider *single-field* inflation, but with a generalized Lagrangian

$$L = P(X,\phi), \qquad X \equiv -\partial_\mu \phi \partial^\mu \phi/2. \tag{194}$$

The category of models, first studied in [64], has been called K-inflation because inflation can arise from the presence of non-standard kinetic terms, and not necessarily from a quasi-flat potential as in standard inflation.

Linear perturbations have been investigated in [65]. Here, one can simply take the single-field limit of (175)

$$S_{(2)} = \frac{1}{2}\int \mathrm{d}t\,\mathrm{d}^3x\,a^3\left[(P_X + 2P_{XX}X)\,\dot{Q}^2 - P_X h^{ij}\partial_i Q\partial_j Q \right. \\ \left. -\mathcal{M}Q^2 + 2\,\Omega Q\dot{Q}\right], \tag{195}$$

where $P_X \equiv \frac{\partial P}{\partial X}$ and $P_X \equiv \frac{\partial^2 P}{\partial X^2}$.
The first line of the above action shows that the perturbations of the scalar field propagate with an effective sound speed given by

$$c_s^2 = \frac{P_X}{P_X + 2XP_{XX}}, \tag{196}$$

which can be much smaller than the usual speed of light, in some models.

Introducing the conformal time τ and the canonically normalized field

$$v = \frac{a\sqrt{P_X}}{c_s}Q, \tag{197}$$

one gets the action

$$S_{(2)} = \frac{1}{2}\int \mathrm{d}\tau\,\mathrm{d}^3x\left[v'^2 - c_s^2(\partial v)^2 + \frac{z''}{z}v^2\right], \tag{198}$$

with

$$z = \frac{a\dot{\phi}\sqrt{P_X}}{c_s H}. \tag{199}$$

In Fourier space, this leads to the equation of motion

$$v'' + \left(k^2c_s^2 - \frac{z''}{z}\right)v = 0, \tag{200}$$

where one notes the presence of c_s^2 multiplying k^2. As a consequence, the fluctuations are amplified at *sound horizon* crossing, i.e. when $kc_s \sim aH$, and not at Hubble radius crossing as in the standard case (the two of course coincide for $c_s \simeq c$).

Assuming a slow variation of the Hubble parameter H and of the sound speed c_s, one can use the approximation $z''/z \simeq 2/\tau^2$ and the solution corresponding to the vacuum on small scales is given by

$$v = \frac{1}{\sqrt{2kc_s}} e^{-ikc_s\tau} \left(1 - \frac{i}{kc_s\tau}\right). \tag{201}$$

This expression differs from (78) only by the presence of c_s.

One can then proceed exactly as in the standard case to obtain the power spectrum of the scalar field fluctuations:

$$\mathcal{P}_Q \simeq \frac{H^2}{4\pi^2 c_s P_X} \tag{202}$$

and the power spectrum of the curvature perturbation:

$$\mathcal{P}_{\mathcal{R}_*} = \frac{k^3}{2\pi^2} \frac{|v_{\sigma\, k}|^2}{z^2} \simeq \frac{H^4}{4\pi^2\dot{\sigma}^2} = \frac{H^2}{8\pi^2\varepsilon c_s}\,, \tag{203}$$

where $\varepsilon = -\dot{H}/H^2$.

6.4 A Specific Example: Multi-field DBI Inflation

The two previous subsections have illustrated *separately* the consequences of multiple inflatons, on the one hand, and of non-standard kinetic terms. Here, these two aspects will be naturally combined in a category of models motivated by string theory.

Indeed, inflation could originate from the motion of a $D3$-brane in an internal six-dimensional compact space. The dynamics of the brane, with tension T_3, is governed by the Dirac-Born-Infeld Lagrangian (we ignore here the dilaton and the various form fields, but they can be included, as in [66])

$$L_{\rm DBI} = -T_3\sqrt{-\det\gamma_{\mu\nu}}, \tag{204}$$

which depends on the determinant of the induced metric on the 3-brane,

$$\gamma_{\mu\nu} = H_{AB}\,\partial_\mu Y^A_{\rm (b)}\partial_\nu Y^B_{\rm (b)}, \tag{205}$$

where H_{AB} is metric of the compactified 10-dimensional spacetime, assumed to be of the form

$$H_{AB}\,dY^A\,dY^B = h^{-1/2}(y^K)\,g_{\mu\nu}dx^\mu dx^\nu + h^{1/2}(y^K)\,G_{IJ}(y^K)\,dy^I dy^J\,, \tag{206}$$

and $Y^A_{\rm (b)}(x^\mu) = (x^\mu, \psi^I(x^\mu))$, with $\mu = 1\ldots3$ and $I = 1\ldots6$, defines the brane embedding.

After using the rescalings $\phi^I \equiv \sqrt{T_3}Y^I$ and $f = h/T_3$, one ends up with a Lagrangian of the form

$$P = -\frac{1}{f(\boldsymbol{\phi}^I)}\left(\sqrt{\mathcal{D}} - 1\right) - V(\boldsymbol{\phi}^I) \tag{207}$$

with

$$\begin{aligned}\mathcal{D} &\equiv \det\left(\delta^\mu_\nu + f\, G_{IJ}\partial^\mu\phi^I\partial_\nu\phi^J\right)\\ &= 1 - 2fG_{IJ}X^{IJ} + 4f^2X^{[I}_I X^{J]}_J - 8f^3X^{[I}_I X^J_J X^{K]}_K + 16f^4X^{[I}_I X^J_J X^K_K X^{L]}_L,\end{aligned} \tag{208}$$

where the field indices are lowered by the field metric G_{IJ}, i.e. the metric of the internal compact space, and the brackets denote antisymmetrization of the indices. A potential term, which arises from the brane's interactions with bulk fields or other branes, is also included.

If the brane moves radially in a conical geometry, one can ignore the angular directions in field space, and the effective action reduces to

$$S = \int d^4x\sqrt{-g}\left[-\frac{1}{f}\left(\sqrt{1 + f\,\partial_\mu\phi\partial^\mu\phi} - 1\right) - V(\phi)\right]. \tag{209}$$

If $f\dot{\phi}^2 \ll 1$, one can expand the square root in the Lagrangian and one recovers the usual kinetic term familiar to slow-roll inflation. But there is another regime, called DBI inflation [67, 68], corresponding to the "relativistic" limit

$$1 - f\,\dot{\phi}^2 \ll 1 \Rightarrow |\dot{\phi}| \simeq 1/\sqrt{f}\,, \tag{210}$$

which does not require a very flat potential as in standard slow-roll inflation.

The Lagrangian in (209) is of the form $P(X,\phi)$, discussed in the previous subsection, with

$$P(X,\phi) = -\frac{1}{f(\phi)}\left(\sqrt{1 - 2fX} - 1\right) - V(\phi), \tag{211}$$

and therefore, using (196),

$$c_s = \sqrt{1 - 2fX} = \frac{1}{P_X}. \tag{212}$$

If the brane is allowed to move in the angular directions, the above single-field simplification is not valid and one must work in a multi-field framework with the Lagrangian (207). The perturbations generated by such a scenario have been studied in detail in [69] and we now summarize the main results.

After decomposing the perturbations into adiabatic and entropy modes, one finds that the single-field results apply to adiabatic mode, so that its spectrum at *sound horizon crossing* is given by

$$\mathcal{P}_{Q_\sigma *} \simeq \frac{H^2}{4\pi^2} \tag{213}$$

(the subscript $*$ here indicates that the corresponding quantity is evaluated at sound horizon crossing $kc_s = aH$).

As for the (canonically normalized) entropy mode, $v_s \equiv (a/\sqrt{c_s})Q_s$, its evolution is governed by the equation

$$v_s'' + \xi v_\sigma' + \left(k^2 c_s^2 - \frac{\alpha''}{\alpha}\right) v_s = 0\,, \qquad \alpha \equiv \frac{a}{\sqrt{c_s}}, \tag{214}$$

where we have neglected a possible coupling with the adiabatic mode and assumed that the effective mass of the entropy mode is small with respect to H. v_s has thus the same spectrum as v_σ, but since the normalization coefficients in front of the adiabatic and entropy modes differ, one finds that the spectrum for the fluctuations along the entropy direction in field space, is given by

$$\mathcal{P}_{Q_s^*} \simeq \frac{H^2}{4\pi^2 c_s^2}, \tag{215}$$

which shows that, for small c_s, the entropic modes are *amplified* with respect to the adiabatic modes:

$$Q_{s*} \simeq \frac{Q_{\sigma *}}{c_s}. \tag{216}$$

Since we are in a multi-field scenario, the curvature perturbation can be modified, after sound horizon crossing, if there is a transfer from the entropic to the adiabatic modes, as we saw earlier. The final curvature perturbation can be formally written as

$$\mathcal{R} = \mathcal{R}_* + T_{\mathcal{RS}}\mathcal{S}_*, \tag{217}$$

where, for convenience, we have introduced the *rescaled* entropy perturbation

$$\mathcal{S} = c_s \frac{H}{\dot\sigma} Q_s\,, \tag{218}$$

defined such that its power spectrum at sound horizon crossing is the same as that of the curvature perturbation, i.e. $\mathcal{P}_{\mathcal{S}_*} = \mathcal{P}_{\mathcal{R}_*}$. The final curvature power spectrum is thus given by

$$\mathcal{P}_\mathcal{R} = (1 + T_{\mathcal{RS}}^2)\mathcal{P}_{\mathcal{R}_*} = \frac{\mathcal{P}_{\mathcal{R}_*}}{\cos^2\Theta}\,, \tag{219}$$

where we have introduced the "transfer angle" Θ ($\Theta = 0$ if there is no transfer and $|\Theta| = \pi/2$ if the final curvature perturbation is mostly of entropic origin) by

$$\sin\Theta = \frac{T_{\mathcal{RS}}}{\sqrt{1+T_{\mathcal{RS}}^2}}\,. \tag{220}$$

The power spectrum for the tensor modes is still governed by the transition at *Hubble radius* and its amplitude, given by (121), is unchanged. The tensor to scalar ratio is thus

$$r \equiv \frac{\mathcal{P}_{\mathcal{T}}}{\mathcal{P}_{\mathcal{R}}} = 16\,\varepsilon\, c_s \cos^2\Theta\,. \tag{221}$$

Interestingly this expression combines the result of k-inflation [65], where the ratio is suppressed by a small sound speed c_s, and that of multi-field inflation with standard kinetic terms [70], where the ratio is suppressed by a large transfer from entropy to adiabatic modes.

6.5 The Curvaton Scenario

The transfer from entropy into adiabatic perturbations can occur during inflation, as we have seen in scenarios with multiple inflatons, but it can also take place long after the end of inflation. A much studied example of this possibility is the curvaton scenario [71–73] (see also [74]).

The curvaton is a weakly coupled scalar field, σ, which is light relative to the Hubble rate during inflation, and hence acquires an almost scale-invariant spectrum and effectively Gaussian distribution of perturbations during inflation:

$$\mathcal{P}_{\delta\sigma} = \left(\frac{H}{2\pi}\right)^2, \tag{222}$$

where the curvaton perturbation, $\delta\sigma = Q_\sigma$, is defined here in the flat gauge.

After inflation the Hubble rate drops and eventually the curvaton becomes non-relativistic so that its energy density grows relative to radiation, until it contributes a significant fraction of the total energy density, $\Omega_\sigma \equiv \bar{\rho}_\sigma/\bar{\rho}$, before it decays. Hence the initial curvaton field perturbations on large scales can give rise to a primordial density perturbation after it decays (Fig. 5).

The non-relativistic curvaton (mass $m \gg H$), before it decays, can be described by a pressureless, non-interacting fluid with energy density

$$\rho_\sigma = m^2\sigma^2\,, \tag{223}$$

where σ is the rms amplitude of the curvaton field, which oscillates on a timescale m^{-1} much less than the Hubble time H^{-1}. The corresponding perturbations are characterized, using (136) and (222),

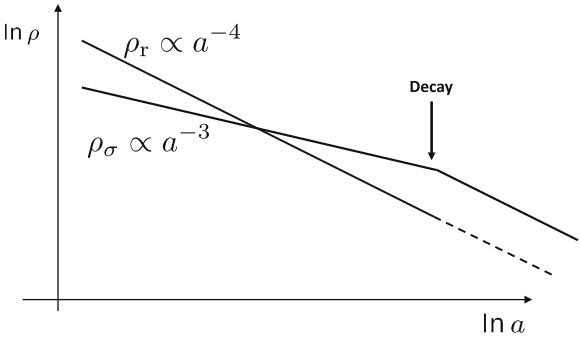

Fig. 5 Evolution of the energy density of the radiation, ρ_r, produced by the inflaton and of the energy density of the curvaton, ρ_σ, before and after the curvaton decay

$$\zeta_\sigma = \left(\frac{\delta\rho_\sigma}{3\rho_\sigma}\right)_{\rm flat} = \frac{2}{3}\frac{\delta\sigma}{\sigma} \quad \Rightarrow \mathcal{P}_{\zeta_\sigma} \simeq \frac{H^2}{9\pi^2\sigma^2}. \tag{224}$$

When the curvaton decays into radiation, its perturbations are converted into perturbations of the resulting radiation fluid. The subsequent perturbation is described by

$$\zeta = r_\sigma \zeta_\sigma + (1 - r_\sigma)\zeta_{\rm inf}\,, \quad r_\sigma \equiv \frac{3\Omega_{\sigma,{\rm decay}}}{4 - \Omega_{\sigma,{\rm decay}}}\,. \tag{225}$$

This implies that the power spectrum for the primordial adiabatic perturbation $\zeta_{\rm r}$ can be expressed as

$$\mathcal{P}_\zeta = \mathcal{P}_{\zeta_{\rm inf}} + r_\sigma^2 \mathcal{P}_{\zeta_\sigma}\,. \tag{226}$$

where $\mathcal{P}_{\zeta_{\rm inf}}$ is the spectrum of perturbations generated directly by the inflaton fluctuations.

In the case of single-field inflation, $\mathcal{P}_{\zeta_{\rm inf}}$ is given in (117) and one can rewrite the total power spectrum as we have

$$\mathcal{P}_\zeta = (1 + \lambda)\mathcal{P}_{\zeta_{\rm inf}}\,, \qquad \lambda \equiv \frac{8}{9} r_\sigma^2 \varepsilon_* \left(\frac{\sigma_*}{M_P}\right)^{-2}. \tag{227}$$

The limit $\lambda \gg 1$ corresponds to the original curvaton scenario where the inflaton perturbations are negligible: since r_σ and ε_* are bounded by 1, this requires $\sigma_* \ll M_P$.

A value of λ of order 1 or smaller is possible if r_σ or ε_* are sufficiently small and/or σ_* is of the order of M_P. In the latter case the curvaton starts to oscillate at about the same time as it decays and cannot be described as a dust field. A more refined treatment [75] shows that the curvature perturbation due to the inflaton and curvaton perturbations is given

$$\mathcal{R} = -\frac{V}{M_P^2 V'}\delta\phi - \frac{3}{2}f(\sigma_*)\frac{\delta\sigma_*}{M_P}, \tag{228}$$

where the function $f(\sigma_*)$ interpolates between the limiting situations of a pure curvaton and of a secondary inflaton

$$f(\sigma_*) \simeq \begin{cases} \frac{4}{9}\frac{M_P}{\sigma_*}, & \sigma_* \ll M_P \\ \\ \frac{\sigma_*}{3M_P}, & \sigma_* \gg M_P. \end{cases} \tag{229}$$

Interestingly, the curvaton scenario can also produce entropic, or isocurvature, perturbations [76]. It can produce a CDM isocurvature perturbation if the CDM is created before the curvaton decay and thus inherits the perturbations of the inflaton so that $S_{\rm cdm} = 3(\zeta_{\rm inf} - \zeta_r)$; or, on the contrary, if the CDM is created by the curvaton decay, in which case $S_{\rm cdm} = 3(\zeta_\chi - \zeta_r)$. Similarly, baryon isocurvature perturbations can be generated if the baryon asymmetry exists before the curvaton decay.

6.6 Modulaton

In the curvaton scenario, the primordial perturbations are mainly due to the curvaton perturbations if the curvaton was the dominant species at some epoch in the past of the Universe. But one can also envisage scenarios where the primordial perturbations are due to the perturbations of a scalar field, which has never dominated the matter content of the Universe but has affected one transition in the cosmological history. We will name this field a *modulaton*.

The best example is the *modulated reheating* scenario [77, 78] where the decay rate of the inflaton, Γ, depends on a modulaton σ, which has acquired classical fluctuations during inflation. From one super-Hubble patch to another, the decay rate is thus slightly different, which generates a curvature perturbation.

A simple way to quantify this effect is to compute the number of e-folds between some initial time t_i during inflation, when the scale of interest crossed out the Hubble radius and some final time t_f. For simplicity, we will assume that, just after the end of inflation, at time t_e, the inflaton behaves like pressureless matter (as is the case for a quadratic potential) until it decays instantaneously at the time t_d characterized by $H_d = \Gamma$. At the decay, the energy density is thus $\rho_d = \rho_e \exp[-3(N_d - N_e)]$ and is transferred into radiation, so that, at time t_f, one gets

$$\rho_f = \rho_d e^{-4(N_f - N_d)} = \rho_e e^{-3(N_f - N_e) - (N_f - N_d)}. \tag{230}$$

Using the relation $\Gamma = H_d = H_f \exp[2(N_f - N_d)]$ to eliminate $(N_f - N_d)$ in (230), we finally obtain

$$N_f = N_e - \frac{1}{3}\ln\frac{\rho_f}{\rho_e} - \frac{1}{6}\ln\frac{\Gamma}{H_f}\,. \tag{231}$$

If one ignores the inflaton fluctuations, the final curvature perturbation is therefore

$$\zeta = N_{,\sigma}\delta\sigma_* = -\frac{1}{6}\frac{\Gamma_{,\sigma}}{\Gamma}\delta\sigma_*, \tag{232}$$

which yields the curvature power spectrum

$$\mathcal{P}_\zeta = \frac{1}{36}\left(\frac{\Gamma_{,\sigma}}{\Gamma}\right)^2\left(\frac{H_*}{2\pi}\right)^2. \tag{233}$$

The dependence on the modulaton can alternatively show up in the mass of the particles created by the decay of the inflaton [79, 80].

The modulaton can also affect the cosmological evolution *during* inflation, as in the *modulated trapping* scenario [81], which is based on the resonant production of particles during inflation [82]. If the inflaton is coupled to some particles, whose effective mass becomes zero for a critical value of the inflaton, then there will be a burst of production of these particles when the inflaton crosses the critical value. These particles will be quickly diluted but they will slow down the inflaton. This effect, which increases the number of e-folds until the end of inflation, can depend on a modulaton, for example, via the coupling between the inflaton and the particles, and a significant curvature perturbation might be generated (see [81] for details).

6.7 "Initial" Adiabatic and Entropic Perturbations

In contrast with single-field inflation, multi-field inflation can generate isocurvature "initial" perturbations in the radiation era. Note that this is only a possibility but not a necessity; purely adiabatic initial conditions are perfectly compatible with multi-field scenarios.

The CMB is a powerful way to study isocurvature perturbations because (primordial) adiabatic and isocurvature perturbations produce very distinctive features in the CMB anisotropies. On large angular scales, one can show, for instance, that [63]

$$\frac{\delta T}{T} \simeq \frac{1}{5}\left(\mathcal{R} - 2S\right). \tag{234}$$

On smaller angular scales, an adiabatic initial perturbation generates a cosine oscillatory mode in the photon-baryon fluid, leading to an acoustic peak at $\ell \simeq 220$ (for a flat Universe), whereas a pure isocurvature initial perturbation generates a sine oscillatory mode resulting in a first peak at $\ell \simeq 330$. The unambiguous observation of the first peak at $\ell \simeq 220$ has eliminated the possibility of a dominant isocurvature

perturbation. The recent observation by WMAP of the CMB polarization has also confirmed that the initial perturbation is mainly an adiabatic mode. But this does not exclude the presence of a subdominant isocurvature contribution, which could be detected in future high-precision experiments such as Planck.

The combined impact of adiabatic and entropic perturbations crucially depends on their correlation [63, 83]

$$\beta = \frac{\mathcal{P}_{S,\mathcal{R}}}{\sqrt{\mathcal{P}_S \mathcal{P}_{\mathcal{R}}}}. \tag{235}$$

Parametrizing the relative amplitude between the two types of perturbations by a coefficient α,

$$\frac{\mathcal{P}_S}{\mathcal{P}_{\mathcal{R}}} \equiv \frac{\alpha}{1-\alpha}\,, \tag{236}$$

the WMAP5 data [84] yield the following constraints on the entropy contribution

$$\beta = 0\text{: } \alpha_0 < 0.067\ (95\%\,\mathrm{CL}) \qquad \beta = -1\text{: } \alpha_{-1} < 0.0037\ (95\%\,\mathrm{CL}) \tag{237}$$

in the uncorrelated case ($\beta = 0$) and in the totally anti-correlated case ($\beta = -1$), respectively.

7 Primordial Non-Gaussianities

One of the most promising probes of the early Universe, which has been investigately very actively in the last few years, is the non-Gaussianity of the primordial perturbations (see [85] for a review, but the field has grown considerably in the last few years). Whereas the simplest models of inflation, based on a *single* field with *standard* kinetic term, produce undetectable levels of non-Gaussianity [35, 86], a significant amount of non-Gaussianity can be produced in scenarios with (i) non-standard kinetic terms; (ii) multiple fields; (iii) a non standard vacuum; (iv) a non slow-roll evolution.

7.1 Higher Order Correlation Functions

The most natural estimate of non-Gaussianity is the bispectrum defined, in Fourier space, by

$$\langle \zeta_{\mathbf{k}_1} \zeta_{\mathbf{k}_2} \zeta_{\mathbf{k}_3} \rangle \equiv (2\pi)^3 \delta^{(3)} \left(\sum_i \mathbf{k}_i \right) B_\zeta(k_1, k_2, k_3)\,, \tag{238}$$

where the Fourier modes are defined by

$$\zeta_{\mathbf{k}} = \int d^3\mathbf{x}\, e^{-i\mathbf{k}\cdot\mathbf{x}}\, \zeta(\mathbf{x})\,. \tag{239}$$

Equivalently, one often uses the so-called $f_{\rm NL}$ parameter, which can be defined in general by

$$B_\zeta(k_1,k_2,k_3) \equiv \frac{6}{5} f_{\rm NL}(k_1,k_2,k_3)\left[P_\zeta(k_1)P_\zeta(k_2) + P_\zeta(k_2)P_\zeta(k_3) + P_\zeta(k_3)P_\zeta(k_1)\right], \tag{240}$$

where P_ζ is the power spectrum[4] defined by

$$\langle \zeta_{\mathbf{k}_1}\zeta_{\mathbf{k}_2}\rangle = (2\pi)^3\delta^{(3)}(\mathbf{k}_1+\mathbf{k}_2)\,P(k_1)\,. \tag{241}$$

The $f_{\rm NL}$ parameter was initially introduced in [87] for a very specific type of non-Gaussianity characterized by

$$\zeta(\mathbf{x}) = \zeta_G(\mathbf{x}) + \frac{3}{5} f_{\rm NL}\zeta_G^2(\mathbf{x})\,, \tag{242}$$

in the physical space, where ζ_G is Gaussian and the factor 3/5 appears because $f_{\rm NL}$ was originally defined with respect to the gravitational potential $\Phi = (3/5)\zeta$, instead of ζ. In this particular case, $f_{\rm NL}$, as defined in (240), is independent of the vectors $\mathbf{k}_i$. But, in general, $f_{\rm NL}$ is a function of the three vectors $\mathbf{k}_i$ (which define a triangle in Fourier space since they are constrained by $\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3 = 0$ as a consequence of homogeneity), and the "shape" of the three-point function is an important characterization of how non-Gaussianity was generated [88].

In the context of multi-field inflation, the so-called δN-formalism [89, 52] is particularly useful to evaluate the primordial non-Gaussianity generated on large scales [90]. The idea is to describe, on scales larger than the Hubble radius, the non-linear evolution of perturbations generated during inflation in terms of the perturbed expansion from an initial hypersurface (usually taken at Hubble crossing during inflation) up to a final uniform-density hypersurface (usually during the radiation-dominated era). Using the Taylor expansion of the number of e-folds given as a function of the initial values of the scalar fields,

$$\zeta \simeq \sum_I N_{,I}\delta\varphi_*^I + \frac{1}{2}\sum_{IJ} N_{,IJ}\delta\varphi_*^I\delta\varphi_*^J \tag{243}$$

one finds [90, 91], in Fourier space,

[4] In this section on non-Gaussianities, we have followed the recent literature and adopted the definition (239) for the Fourier modes, which differs slightly from our convention (69) of the previous chapters. This changes the expression of the power spectrum, but the quantity $\mathcal{P}(k)$ is the same in the two conventions.

$$\langle \zeta_{\mathbf{k}_1} \zeta_{\mathbf{k}_2} \zeta_{\mathbf{k}_3} \rangle = \sum_{IJK} N_{,I} N_{,J} N_{,K} \langle \delta\varphi^I_{\mathbf{k}_1} \delta\varphi^J_{\mathbf{k}_2} \delta\varphi^K_{\mathbf{k}_3} \rangle + \frac{1}{2} \sum_{IJKL} N_{,I} N_{,J} N_{,KL} \langle \delta\varphi^I_{\mathbf{k}_1} \delta\varphi^J_{\mathbf{k}_2} (\delta\varphi^K \star \delta\varphi^L)_{\mathbf{k}_3} \rangle + \text{perms.} \tag{244}$$

As illustrated by the first line of the above expression, one sees that significant non-Gaussianities can arise from the three-point function of the scalar field(s). This is the case for models with non-standard kinetic terms [92–94], leading to a specific shape of non-Gaussianities, usually called *equilateral*, where the dominant contribution comes from configurations where the three wavevectors have similar length $k_1 \sim k_2 \sim k_3$. With a non-standard vacuum for the perturbations, one finds a *folded* shape, which is peaked in the limit $k_3 \sim k_1 + k_2$ [94].

Even if the three-point function of the scalar fields is negligible, the second line can lead to sizable non-Gaussianities. Indeed substituting

$$\langle \delta\varphi^I_{\mathbf{k}_1} \delta\varphi^J_{\mathbf{k}_2} \rangle = (2\pi)^3 \delta_{IJ} \delta^{(3)}(\mathbf{k}_1 + \mathbf{k}_2) \frac{2\pi^2}{k_1^3} \mathcal{P}_*(k_1), \qquad \mathcal{P}_*(k) \equiv \frac{H_*^2}{4\pi^2}, \tag{245}$$

in (244), one gets

$$\frac{6}{5} f_{\rm NL} = \frac{N_I N_J N^{IJ}}{(N_K N^K)^2} \,. \tag{246}$$

This corresponds to another type of non-Gaussianity, usually called *local* or *squeezed*, for which the dominant contribution comes from configurations where the three wavevectors form a squeezed triangle.

The present observational constraints [84] are

$$-9 < f_{NL}^{(\rm local)} < 111 \quad (95\%\,\text{CL}), \qquad -151 < f_{NL}^{(\rm equil)} < 253 \quad (95\%\,\text{CL}), \tag{247}$$

for the local non-linear coupling parameter and the equilateral non-linear coupling parameter, respectively.

Extending the Taylor expansion (243) up to third order, one can compute in a similar way the trispectrum [95], i.e. the Fourier transform of the connected four-point function defined by

$$\langle \zeta_{\mathbf{k}_1} \zeta_{\mathbf{k}_2} \zeta_{\mathbf{k}_3} \zeta_{\mathbf{k}_4} \rangle_c \equiv (2\pi)^3 \delta^{(3)}\Big(\sum_i \mathbf{k}_i\Big) T_\zeta(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3, \mathbf{k}_4) \,. \tag{248}$$

Assuming the scalar field perturbations to be quasi-Gaussian, the trispectrum can be written in the form [96]

$$T_\zeta(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3,\mathbf{k}_4) = \tau_{\rm NL}\left[P(k_{13})P(k_3)P(k_4) + 11 \text{ perms}\right] \tag{249}$$

$$+\frac{54}{25}g_{\rm NL}\left[P(k_2)P(k_3)P(k_4) + 3 \text{ perms}\right], \tag{250}$$

with

$$\tau_{\rm NL} = \frac{N_{IJ}N^{IK}N^J N_K}{(N_L N^L)^3}, \qquad g_{\rm NL} = \frac{25}{54}\frac{N_{IJK}N^I N^J N^K}{(N_L N^L)^3} \tag{251}$$

and where $k_{13} \equiv |\mathbf{k}_1 + \mathbf{k}_3|$.

7.2 A Few Examples

It is not always easy to obtain significant non-Gaussianities even in models with several inflatons (see, e.g. [97]). But detectable local non-Gaussianity can be generated in the curvaton or modulaton scenarios, or at the end of inflation [98, 99]. Models with non-standard kinetic terms, like the DBI scenario, can produce strong equilateral non-Gaussianities. We now discuss the curvaton case and multi-field DBI inflation.

7.2.1 Curvaton

In the multi-field scenarios with a curvaton (or a modulaton), one can separate the contributions of the inflaton field ϕ and of the curvaton/modulaton σ. In the case of standard slow-roll inflation, the second derivatives with respect to ϕ are negligible in (246) and one gets

$$\frac{6}{5}f_{\rm NL} = \frac{N_\sigma^2 N_{\sigma\sigma}}{(N_\phi^2 + N_\sigma^2)^2} = \frac{N_{\sigma\sigma}}{N_\sigma^2(1+\lambda^{-1})^2}, \tag{252}$$

where we have introduced the parameter $\lambda \equiv N_\sigma^2/N_\phi^2$, which represents the ratio of the contribution of σ with the inflaton contribution in the power spectrum (see (227) for the curvaton).

For the curvaton, (225) tells us that $N_\sigma = 2r_\sigma/3\sigma$ and the extension of this equation to second order yields

$$N_{\sigma\sigma} = \frac{4r_\sigma}{9\sigma^2}\left(\frac{3}{2} - 2r_\sigma - r_\sigma^2\right). \tag{253}$$

This leads to a local non-Gaussianity characterized by

$$\frac{6}{5}f_{\rm NL} = \frac{1}{r_\sigma}\frac{\left(\frac{3}{2} - 2r_\sigma - r_\sigma^2\right)}{(1+\lambda^{-1})^2}. \tag{254}$$

Non-Gaussianities are thus significant when the curvaton decays well before it dominates, $r_\sigma \ll 1$. When $\lambda \gg 1$ and the perturbations from inflation are negligible, one recovers the standard curvaton result [100].

Note, however, that $f_{\rm NL}$ does not grow indefinitely as r_σ becomes small because both r_σ and λ depend on the curvaton expectation value σ_*. Indeed, substituting $r_\sigma \sim (\sigma_*/M_P)^2/\sqrt{\Gamma_\sigma/m_\sigma}$ (valid in the limit $r \ll 1$), where Γ_σ is the decay rate of the curvaton, into the definition (227), one sees that λ is proportional to σ_*^2, like r. One thus finds [101] that the non-linearity parameter reaches its maximal value $f_{\rm NL}(\,\max\,) \sim \varepsilon_*/\sqrt{\Gamma_\sigma/m_\sigma}$ for $\lambda \sim 1$, i.e. for $\sigma_* \sim \sqrt{\Gamma_\sigma/(m_\sigma \varepsilon_*)} M_P$. A significant non-Gaussianity is thus possible if $\varepsilon_* \gg \sqrt{\Gamma_\sigma/m_\sigma}$. It is easy to extend the above procedure for the computation of the trispectrum [102].

Moreover, in the curvaton scenarios, isocurvature perturbations can be present. Even if their contribution to the power spectrum is constrained to be small, they could contribute significantly to non-Gaussianities. It is thus interesting to study the non-Gaussianities of isocurvature perturbations as well (see, e.g. [103–106]).

Non-Gaussianities in modulaton scenarios have also been investigated (see e.g., [107, 108, 81]).

7.2.2 Multi-field DBI Inflation

Multi-field DBI inflation is another multi-field example where non-Gaussianities have been investigated. In this case, the three-point correlation functions of the scalar fields are not negligible and they can be computed from the third-order action, which is given, in the small sound speed limit, by [69, 51]

$$S_{(3)} = \int \mathrm{d}t\, \mathrm{d}^3x \left\{ \frac{a^3}{2c_s^5\dot{\sigma}} \left[(\dot{Q}_\sigma)^3 + c_s^2 \dot{Q}_\sigma (\dot{Q}_s)^2 \right] - \frac{a}{2c_s^3\dot{\sigma}} \left[\dot{Q}_\sigma (\partial Q_\sigma)^2 - c_s^2 \dot{Q}_\sigma (\partial Q_s)^2 + 2c_s^2 \dot{Q}_s \partial Q_\sigma \partial Q_s \right] \right\}, \qquad (255)$$

in terms of the instantaneous adiabatic and entropic perturbations, respectively. The contribution from the scalar field three-point functions to the coefficient $f_{\rm NL}$ is

$$f_{NL}^{(3)} = -\frac{35}{108}\frac{1}{c_s^2}\frac{1}{1+T_{\mathcal{RS}}^2} = -\frac{35}{108}\frac{1}{c_s^2}\cos^2\Theta, \qquad (256)$$

which is similar to the single-field DBI result, but with a suppression due to the transfer between the entropic and adiabatic modes.

Interestingly, multi-field DBI inflation could also produce a local non-Gaussianity in addition to the equilateral one (see [109] for an explicit illustration). The trispectrum in multi-field DBI inflation has also been computed [110].

8 Conclusions

As these notes have tried to emphasize, inflation provides an attractive framework to describe the very early Universe and to account for the "initial" seeds of the cosmological perturbations, which we are able to observe today with increasing precision. In particular, the idea that the present structures in the Universe arose from the gravitational amplification of quantum vacuum fluctuations is especially appealing.

At present, inflation is more a general framework than a specific theory and there exists a plethora of models, based on various types of motivation, which can all satisfy the present observational data. The simplest models, based on a slow-rolling single field, produce only adiabatic perturbations, with negligible non-Gaussianities, but with a possibly detectable amount of gravitational waves for the large-field subclass.

More sophisticated models, involving multiple scalar fields or non-standard kinetic terms, can lead to a much richer spectrum of possibilities: isocurvature perturbations that could be correlated with the adiabatic ones or a detectable level of non-Gaussianities.

Any clear evidence in the future of one or several of these additional features (gravitional waves, isocurvature perturbations and/or primordial non-Gaussianities) would allow us to discriminate between the main species of inflationary models and would thus have a huge impact on our understanding of the early Universe.

References

1. A. Linde, *Particle Physics and Inflationary Cosmology*, (Harwood, Chur, 1990). 2
2. A.R. Liddle and D.H. Lyth, *Cosmological Inflation and Large-Scale Structure*, (Cambridge University Press, Cambridge, 2000). 2, 10
3. V. Mukhanov, *Physical Foundations of Cosmology*, (Cambridge University Press, Cambridge 2005). 2
4. J.E. Lidsey, A.R. Liddle, E.W. Kolb, E.J. Copeland, T. Barreiro and M. Abney, Rev. Mod. Phys. **69**, 373 (1997) [arXiv:astro-ph/9508078]. 2
5. D.H. Lyth and A. Riotto, Phys. Rept. **314**, 1 (1998) [hep-ph/9807278]. 2
6. A. Riotto, arXiv:hep-ph/0210162. 2
7. D. Langlois, "Inflation, quantum fluctuations and cosmological perturbations," in *Cargese 2003, Particle Physics and Cosmology*, pp. 235–278 [arXiv:hep-th/0405053]. 2, 19
8. B.A. Bassett, S. Tsujikawa and D. Wands, Rev. Mod. Phys. **78**, 537 (2006) [arXiv:astro-ph/0507632]. 2
9. D. Langlois, Prog. Theor. Phys. Suppl. **163**, 258 (2006) [arXiv:hep-th/0509231]. 4
10. A.R. Liddle and S.M. Leach, Phys. Rev. D **68**, 103503 (2003) [arXiv:astro-ph/0305263]. 10
11. A.D. Linde, Phys. Rev. D **49**, 748 (1994) [arXiv:astro-ph/9307002]. 13
12. A.H. Guth, Phys. Rev. D **23**, 347 (1981). 13
13. A.A. Starobinsky, Phys. Lett. B **91**, 99 (1980). 13
14. A.D. Linde, Phys. Lett. B **108**, 389 (1982). 13
15. A. Albrecht and P. J. Steinhardt, Phys. Rev. Lett. **48**, 1220 (1982). 13
16. A.D. Linde, Phys. Lett. B **129**, 177 (1983). 14
17. L. McAllister and E. Silverstein, Gen. Rel. Grav. **40**, 565 (2008) [arXiv:0710.2951 [hep-th]]. 14

18. C.P. Burgess, PoS **P2GC** (2006) 008 [Class. Quant. Grav. **24**, S795 (2007)] [arXiv:0708.2865 [hep-th]]. 14
19. R. Kallosh, *On Inflation in String Theory*. Lect. Notes Phys. **738** 119 (2008) [arXiv:hep-th/0702059]. 14
20. J.M. Cline, arXiv:hep-th/0612129. 14
21. S.H. Henry Tye, *Brane Inflation: String Theory Viewed from the Cosmos*. Lect. Notes Phys. **737**, 949 (2008) [arXiv:hep-th/0610221]. 14
22. V.F. Mukhanov and G. V. Chibisov, JETP Lett. **33**, 532 (1981) [Pisma Zh. Eksp. Teor. Fiz. **33**, 549 (1981)]. 15
23. A.H. Guth and S. Y. Pi, Phys. Rev. Lett. **49**, 1110 (1982). 15
24. A.A. Starobinsky, Phys. Lett. B **117**, 175 (1982). 15
25. S.W. Hawking, Phys. Lett. B **115**, 295 (1982). 15
26. V.F. Mukhanov, JETP Lett. **41**, 493 (1985) [Pisma Zh. Eksp. Teor. Fiz. **41**, 402 (1985)]. 15, 22
27. V. F. Mukhanov, Sov. Phys. JETP **67**, 1297 (1988) [Zh. Eksp. Teor. Fiz. **94N7**, 1 (1988)]. 15
28. J.M. Bardeen, Phys. Rev. D **22**, 1882 (1980). 19
29. H. Kodama and M. Sasaki, Prog. Theor. Phys. Suppl. **78**, 1 (1984). 19
30. V.F. Mukhanov, H. A. Feldman and R. H. Brandenberger, Phys. Rept. **215**, 203 (1992). 19, 22
31. K.A. Malik and D. Wands, Phys. Rept. **475**, 1 (2009) [arXiv:0809.4944 [astro-ph]]. 19
32. M. Sasaki, Prog. Theor. Phys. **76**, 1036 (1986). 22
33. D. Langlois, Class. Quant. Grav. **11**, 389 (1994). 22
34. S. Anderegg and V. F. Mukhanov, Phys. Lett. B **331**, 30 (1994) [arXiv:hep-th/9403091]. 22
35. J.M. Maldacena, JHEP **0305**, 013 (2003) [arXiv:astro-ph/0210603]. 22, 49
36. R. Arnowitt, S. Deser, C.W. Misner, in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962) [arXiv:gr-qc/0405109] 22
37. A.A. Starobinsky, JETP Lett. **30**, 682 (1979) [Pisma Zh. Eksp. Teor. Fiz. **30**, 719 (1979)]. 26
38. G.F.R. Ellis and M. Bruni, Phys. Rev. D **40**, 1804 (1989). 27
39. D. Langlois and F. Vernizzi, Phys. Rev. Lett. **95**, 091303 (2005) [arXiv:astro-ph/0503416]. 28
40. D. Langlois and F. Vernizzi, Phys. Rev. D **72**, 103501 (2005) [arXiv:astro-ph/0509078]. 28
41. D. Langlois and F. Vernizzi, JCAP **0602**, 014 (2006) [arXiv:astro-ph/0601271]. 28
42. D.H. Lyth, K. A. Malik and M. Sasaki, JCAP **0505**, 004 (2005) [arXiv:astro-ph/0411220]. 28
43. G.I. Rigopoulos and E. P. S. Shellard, Phys. Rev. D **68**, 123518 (2003) [arXiv:astro-ph/0306620]. 28
44. J.M. Bardeen, P. J. Steinhardt and M. S. Turner, Phys. Rev. D **28**, 679 (1983). 29
45. D. Wands, K. A. Malik, D. H. Lyth and A. R. Liddle, Phys. Rev. D **62**, 043527 (2000) [arXiv:astro-ph/0003278]. 29
46. J. Martin and R. H. Brandenberger, Phys. Rev. D **63**, 123501 (2001) [arXiv:hep-th/0005209]. 30
47. S. Weinberg, Phys. Rev. D **70**, 043541 (2004) [arXiv:astro-ph/0401313]. 31
48. R. Durrer, *The Cosmic Microwave Background*, (Cambridge University Press, Cambridge, UK, 2008). 32
49. A. Challinor and H. Peiris, AIP Conf. Proc. **1132**, 86 (2009) [arXiv:0903.5158 [astro-ph.CO]]. 32
50. D.H. Lyth, Phys. Rev. Lett. **78**, 1861 (1997) [arXiv:hep-ph/9606387]. 34
51. D. Langlois, S. Renaux-Petel, D. A. Steer and T. Tanaka, Phys. Rev. D **78**, 063523 (2008) [arXiv:0806.0336 [hep-th]]. 36, 53
52. M. Sasaki and E.D. Stewart, Prog. Theor. Phys. **95**, 71 (1996) [arXiv:astro-ph/9507001]; 37, 50
53. S. Groot Nibbelink and B. J. W. van Tent, Class. Quant. Grav. **19**, 613 (2002) [arXiv:hep-ph/0107272]. 37
54. D. Langlois and S. Renaux-Petel, JCAP **0804**, 017 (2008) [arXiv:0801.1085 [hep-th]]. 37, 38
55. C. Gordon, D. Wands, B. A. Bassett and R. Maartens, Phys. Rev. D **63**, 023506 (2001) [arXiv:astro-ph/0009131]. 37
56. F. Di Marco, F. Finelli and R. Brandenberger, Phys. Rev. D **67**, 063512 (2003) [arXiv:astro-ph/0211276]. 39
57. Z. Lalak, D. Langlois, S. Pokorski and K. Turzynski, JCAP **0707**, 014 (2007) [arXiv:0704.0212 [hep-th]]. 39, 40

58. D. Langlois and F. Vernizzi, JCAP **0702**, 017 (2007) [arXiv:astro-ph/0610064]. 39
59. S. Renaux-Petel and G. Tasinato, JCAP **0901**, 012 (2009) [arXiv:0810.2405 [hep-th]]. 39
60. G.I. Rigopoulos, E.P.S. Shellard and B.J.W. van Tent, Phys. Rev. D **73**, 083521 (2006) [arXiv:astro-ph/0504508]. 39
61. M. Sasaki and T. Tanaka, Prog. Theor. Phys. **99**, 763 (1998) [arXiv:gr-qc/9801017]. 39
62. A.A. Starobinsky and J. Yokoyama, "Density fluctuations in Brans-Dicke inflation", gr-qc/9502002 40
63. D. Langlois, Phys. Rev. D **59**, 123512 (1999) [arXiv:astro-ph/9906080]. 40, 48, 49
64. C. Armendariz-Picon, T. Damour and V. F. Mukhanov, Phys. Lett. B **458**, 209 (1999) [arXiv:hep-th/9904075]. 41
65. J. Garriga and V.F. Mukhanov, Phys. Lett. B **458**, 219 (1999) [arXiv:hep-th/9904176]. 41, 45
66. D. Langlois, S. Renaux-Petel and D.A. Steer, JCAP **0904**, 021 (2009) [arXiv:0902.2941 [hep-th]]. 42
67. E. Silverstein and D. Tong, Phys. Rev. D **70**, 103505 (2004) [arXiv:hep-th/0310221]; 43
68. M. Alishahiha, E. Silverstein and D. Tong, Phys. Rev. D **70**, 123505 (2004) [arXiv:hep-th/0404084]. 43
69. D. Langlois, S. Renaux-Petel, D. A. Steer and T. Tanaka, Phys. Rev. Lett. **101**, 061301 (2008) [arXiv:0804.3139 [hep-th]]. 43, 53
70. N. Bartolo, S. Matarrese and A. Riotto, Phys. Rev. D **64**, 123504 (2001) [arXiv:astro-ph/0107502]. 45
71. K. Enqvist and M. S. Sloth, Nucl. Phys. B **626**, 395 (2002) [arXiv:hep-ph/0109214]; 45
72. D. H. Lyth and D. Wands, Phys. Lett. B **524**, 5 (2002) [arXiv:hep-ph/0110002]; 45
73. T. Moroi and T. Takahashi, Phys. Lett. B **522**, 215 (2001) [Erratum-ibid. B **539**, 303 (2002)] [arXiv:hep-ph/0110096]. 45
74. A.D. Linde and V.F. Mukhanov, Phys. Rev. D **56**, 535 (1997) [arXiv:astro-ph/9610219]. 45
75. D. Langlois and F. Vernizzi, Phys. Rev. D **70**, 063522 (2004) [arXiv:astro-ph/0403258]. 46
76. D.H. Lyth, C. Ungarelli and D. Wands, Phys. Rev. D **67**, 023503 (2003) [arXiv:astro-ph/0208055]. 47
77. G. Dvali, A. Gruzinov and M. Zaldarriaga, Phys. Rev. D **69**, 023505 (2004) [arXiv:astro-ph/0303591]. 47
78. L. Kofman, arXiv:astro-ph/0303614. 47
79. G. Dvali, A. Gruzinov and M. Zaldarriaga, Phys. Rev. D **69**, 083505 (2004) [arXiv:astro-ph/0305548]. 48
80. F. Vernizzi, Phys. Rev. D **69**, 083526 (2004) [arXiv:astro-ph/0311167]. 48
81. D. Langlois and L. Sorbo, arXiv:0906.1813 [astro-ph.CO]. 48, 53
82. D. J. H. Chung, E. W. Kolb, A. Riotto and I. I. Tkachev, Phys. Rev. D **62**, 043508 (2000) [arXiv:hep-ph/9910437]. 48
83. D. Langlois and A. Riazuelo, Phys. Rev. D **62**, 043504 (2000) [arXiv:astro-ph/9912497]. 49
84. E. Komatsu et al. [WMAP Collaboration], arXiv:0803.0547 [astro-ph]. 49, 51
85. N. Bartolo, E. Komatsu, S. Matarrese and A. Riotto, Phys. Rept. **402**, 103 (2004) [arXiv:astro-ph/0406398]. 49
86. V. Acquaviva, N. Bartolo, S. Matarrese and A. Riotto, Nucl. Phys. B **667**, 119 (2003) [arXiv:astro-ph/0209156]. 49
87. E. Komatsu and D.N. Spergel, Phys. Rev. D **63**, 063002 (2001) [arXiv:astro-ph/0005036]. 50
88. D. Babich, P. Creminelli and M. Zaldarriaga, JCAP **0408**, 009 (2004) [arXiv:astro-ph/0405356]. 50
89. A. A. Starobinsky, JETP Lett. **42**, 152 (1985) [Pisma Zh. Eksp. Teor. Fiz. **42**, 124 (1985)]. 50
90. D. H. Lyth and Y. Rodriguez, Phys. Rev. Lett. **95**, 121302 (2005) [arXiv:astro-ph/0504045]. 50
91. D. Seery and J.E. Lidsey, JCAP **0509**, 011 (2005) [arXiv:astro-ph/0506056]. 50
92. P. Creminelli, JCAP **0310**, 003 (2003) [arXiv:astro-ph/0306122]. 51
93. D. Seery and J.E. Lidsey, JCAP **0506**, 003 (2005) [arXiv:astro-ph/0503692]. 51
94. X. Chen, M. x. Huang, S. Kachru and G. Shiu, JCAP **0701**, 002 (2007) [arXiv:hep-th/0605045]. 51

95. D. Seery, J.E. Lidsey and M.S. Sloth, JCAP **0701**, 027 (2007) [arXiv:astro-ph/0610210]. 51
96. C.T. Byrnes, M. Sasaki and D. Wands, Phys. Rev. D **74**, 123519 (2006) [arXiv:astro-ph/0611075]. 51
97. F. Vernizzi and D. Wands, JCAP **0605**, 019 (2006) [arXiv:astro-ph/0603799]. 52
98. F. Bernardeau and J.P. Uzan, Phys. Rev. D **67**, 121301 (2003) [arXiv:astro-ph/0209330]. 52
99. M. Sasaki, Prog. Theor. Phys. **120**, 159 (2008) [arXiv:0805.0974 [astro-ph]]. 52
100. N. Bartolo, S. Matarrese and A. Riotto, JCAP **0401**, 003 (2004) [arXiv:astro-ph/0309692]. 53
101. K. Ichikawa, T. Suyama, T. Takahashi and M. Yamaguchi, Phys. Rev. D **78**, 023513 (2008) [arXiv:0802.4138 [astro-ph]]. 53
102. M. Sasaki, J. Valiviita and D. Wands, Phys. Rev. D **74**, 103003 (2006) [arXiv:astro-ph/0607627]. 53
103. M. Kawasaki, K. Nakayama, T. Sekiguchi, T. Suyama and F. Takahashi, JCAP **0811**, 019 (2008) [arXiv:0808.0009 [astro-ph]]. 53
104. D. Langlois, F. Vernizzi and D. Wands, JCAP **0812**, 004 (2008) [arXiv:0809.4646 [astro-ph]]. 53
105. M. Kawasaki, K. Nakayama, T. Sekiguchi, T. Suyama and F. Takahashi, JCAP **0901**, 042 (2009) [arXiv:0810.0208 [astro-ph]]. 53
106. T. Moroi and T. Takahashi, Phys. Lett. B **671**, 339 (2009) [arXiv:0810.0189 [hep-ph]]. 53
107. M. Zaldarriaga, Phys. Rev. D **69**, 043508 (2004) [arXiv:astro-ph/0306006]. 53
108. K. Ichikawa, T. Suyama, T. Takahashi and M. Yamaguchi, Phys. Rev. D **78**, 063545 (2008) [arXiv:0807.3988 [astro-ph]]. 53
109. S. Renaux-Petel, arXiv:0907.2476 [hep-th]. 53
110. S. Mizuno, F. Arroja and K. Koyama, arXiv:0907.2439 [hep-th]. 53

Type Ia Supernovae and Cosmology

M. Sullivan

Summary Type Ia Supernovae (SNe Ia) provide cosmologists with a precise calibrateable "standard candle" with which to probe the expansion history of the universe on large scales. Pioneering astronomical surveys in the late 1990s exploited these distant cosmic explosions to directly reveal the presence of a "dark energy," opposing the attractive, slowing force of gravity and instead accelerating the universe's rate of expansion. Dark energy has since emerged as being responsible for more than 70% of the universe's mass–energy: the lack of a viable theoretical explanation has sparked an intense observational effort to understand its nature. In this chapter we review the use of SNe Ia in cosmology and dark energy studies. We begin by placing the SNe Ia in a cosmological context, introducing the framework in which their physical fluxes are interpreted, and discussing their underlying physics which leads to their near-uniform peak brightness, exploited by astronomers to estimate distances. We show how advances in the empirical understanding of SNe Ia led to the direct discovery of the accelerating universe and how modern SN Ia searches and distance estimation techniques, combined with complementary probes of large-scale structure such as baryon acoustic oscillations, have measured the average equation of state of dark energy to better than 5% (statistical error). Systematics are now of increasing importance and we discuss the main sources of these, both experimental and astrophysical, together with an experimental error budget typical of that in a modern SN Ia survey. Finally, we outline the future prospects for measuring dark energy with SN Ia using the next generation of planned experiments.

1 Introduction

The newest puzzle in cosmology is the observed acceleration of the expansion of the universe. The universe has been known to be expanding, and not static and unchanging, since the beginning of the twentieth century. The early work of Slipher, Hubble,

M. Sullivan (✉)
Department of Physics (Astrophysics), DWB, Keble Road, Oxford, OX1 3RH, UK
e-mail: sullivan@astro.ox.ac.uk

Sullivan, M.: *Type Ia Supernovae and Cosmology*. Lect. Notes Phys. **800**, 59–97 (2010)
DOI 10.1007/978-3-642-10598-2_2

and Humason [1, 2] showed that nearby "spiral nebulae" are receding from the Earth in every direction on the sky with velocities proportional to their inferred distance, implying that the universe is getting larger over time, expanding in every direction. For a universe filled with matter and radiation, general relativity (GR) predicts that the gravitational attraction of the matter in the universe should lead to a *deceleration* in this expansion rate as the universe grows and ages. However, observations over the last decade have shown the exact opposite: the rate of the expansion is increasing with time; the expansion of the universe is *accelerating*. This "cosmic acceleration" has been confirmed with a wide variety of different astrophysical observations, and the data indicating this acceleration are now not seriously in question. *However, the physical reason for the observed cosmic acceleration remains a complete mystery.*

Two broad possibilities are generally considered. The first is that around 70% of the matter–energy density of the universe exists in an as yet unknown form, coined "dark energy," the key characteristic of which is a strong negative pressure which pushes the universe apart. There exists no compelling or elegant explanation for the presence or nature of this dark energy, or the magnitude of its observed influence, although various theoretical possibilities have been postulated [e.g., 3, 4]. Dark energy could be in the form of a vacuum energy, filling the universe and constant in space and time – a "Cosmological Constant." Alternatively, dark energy may be dynamical, a rolling scalar energy field which varies with both time and location ("quintessence" theories).

A second possibility is that the observed cosmic acceleration is an artifact of our incomplete knowledge of physical laws of gravity in the universe, in particular that the laws of GR, a foundation of modern physics, simply break down on the largest scales. The implication of this is that the cosmological framework in which we interpret astronomical observations is incorrect, and this is manifested in observational data as an acceleration in the expansion rate. These ideas are collectively known as "modified gravity" theories. Such theories are constrained in that they must be essentially equivalent to GR on scales of the solar system, where GR is stunningly successful, and also in the early universe where the predictions of standard cosmology match observational effects such as the properties of the cosmic microwave background and the growth of large-scale structure [5]. Nonetheless, a confirmation of this alternative explanation for the observed acceleration would be as profound as the existence of dark energy itself.

Either of these possibilities would revolutionize our understanding of the laws governing the physical evolution of the universe. Understanding the cosmic acceleration has, therefore, rapidly developed over the last decade into a key goal of modern science [6–9]. This chapter is aimed at the observational part of this effort and, in particular, the use of Type Ia Supernovae (SNe Ia) to probe the expansion history. (We do not tackle the theoretical possibilities for explaining the cosmic acceleration in any great detail; excellent reviews of these can be found elsewhere [e.g., 4].)

Type Ia Supernovae (SNe Ia) are a violent endpoint of stellar evolution, the result of the thermonuclear destruction of an accreting carbon–oxygen white dwarf star approaching the Chandrasekhar mass limit, the maximum theoretical mass that a

white dwarf star can attain before the electron degeneracy pressure supporting it against gravitational collapse, is no longer of sufficient strength. As the white dwarf star gains material from a binary companion and approaches this mass limit, the core temperature of the star increases leading to a runaway fusion of the nuclei in the white dwarf's interior. The kinetic energy release from this nuclear burning – some 10^{44} J – is sufficient to dramatically unbind the star. The resulting violent explosion and shock wave appears billions of times brighter than our Sun, comfortably outshining the galaxy in which the white dwarf resided.

Remarkably, SNe Ia are also extraordinary examples of a class of objects known as standard candles, objects with a uniform intrinsic brightness. For SNe Ia, this homogeneity is presumably due to the similarity of the triggering white dwarf mass (i.e., the Chandrasekhar mass, $\sim$1.4 solar masses or $M_{\odot}$) and consequently the amount of nuclear fuel available to burn. This makes SNe Ia the best (or at least most practical) example of "standard candles" in the distant universe, objects to which a distance can be inferred from only a measurement of the apparent brightness on the sky. This allows them to be used to directly trace the expansion rate of the universe.

In this chapter, I will introduce the framework within which SN Ia observations can be interpreted and show how they can be used to constrain the cosmic expansion history. In particular, I will concentrate on the potential systematic issues that could affect their use and show how future cosmological surveys are being designed to mitigate these effects.

2 Context and Basic Concepts

The key problem with understanding dark energy is its very low density – less than 10^{-29} gcm^{-3} – which makes detecting it in a laboratory, let alone studying it in detail, currently impossible. The only reason that dark energy has such an important measurable effect on the physical evolution of the universe is that it is thought to uniformly fill the cosmos. When this low density is integrated over cosmological distances, its effect dominates over that of matter, which is extremely clustered in stars and galaxies. The influence of dark energy can therefore only be observed over cosmological scales, which in turn makes astronomy the only experimental field currently capable of making headway in studying it.

Although there has been much excitement over the last decade, cosmic acceleration is neither a new nor novel concept. Its history can be traced back to the development of the theory of GR, and the idea has re-emerged several times in the intervening century [for a "pre-1998" review see 10]. At the time of the publication of GR, contemporary thinking indicated that the universe was a static place. Einstein perceived that solutions to the field equations of GR did not allow for these static solutions where space is neither expanding nor contracting but rather is dynamically stable. The effects of gravity in any universe containing matter would cause that universe to eventually collapse. Hence, Einstein famously added a repulsive "cosmological constant" term to his equations – Λ.

This cosmological constant has the same effect mathematically as an intrinsic energy density of the vacuum with an associated pressure. A positive vacuum energy density implies a negative pressure (and vice versa). If the vacuum energy density is positive, this negative pressure will drive an accelerated expansion of empty space, acting against the slowing force of gravity. Hence, static universe solutions in GR could now be permitted, at least in principle. Following observations in the late 1920s that the universe was not a static place but instead expands with time, the perceived need for a Λ term in GR was removed. Einstein famously remarked in his later life that modifying his original equations of GR to include Λ was his "biggest blunder."

Despite Einstein's retraction of Λ, in the early 1990s it was realized that, in fact, a cosmological constant could potentially explain many puzzling observational effects in astronomical data. Many cosmologists were disturbed by the low matter density implied by observations of the large-scale structure of the universe – if $\Omega_{\mathrm{M}} < 1$, where was the rest of the matter–energy? Was the universe non-flat or was the interpretation of the observations at fault? The apparent ages of globular clusters were another puzzle, seemingly older than the accepted age of the universe in the then standard cosmological models. This indirect evidence generated a renewed interest in the cosmological constant, which could explain many of these inconsistencies [e.g., 11–13]. However, the first direct evidence did not come until a few years later with observations of SNe Ia.

2.1 Cosmological Framework

In the solutions to Einstein's field equations of GR known as the Friedmann–Lemaitre–Robertson–Walker (FLRW) metric, the universe is described as homogeneous and isotropic, possibly expanding or contracting, and filled with a perfect fluid, one that can be completely characterized by an equation of state w, with an energy density ρ and an isotropic pressure p ($w = p/\rho$). In these solutions, the growth of the universe over time is parametrized by a dimensionless scale factor parameter $a(t)$, essentially describing how the universe "stretches" over time, defined so that at the present day $a = 1$. The equations which govern the expansion are known as the Friedmann equations

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho\left(a\right) + 3p(a)\right) \tag{1}$$

and

$$\left(\frac{\dot{a}}{a}\right)^2 \equiv H^2\left(a\right) = \frac{8\pi G\rho\left(a\right)}{3} - \frac{k}{a^2}\,. \tag{2}$$

The left-hand side of this second equation is the Hubble parameter $H(a)$, which measures the relative expansion rate of the universe as a function of a. Despite a

contentious history, the present-day value of H, the Hubble constant H_0, is generally agreed to be close to $70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ using a wide variety of different techniques [e.g., 14]. The right-hand side of (2) determines the expansion rate from the matter–energy contents of the universe. $\rho(a)$ describes the evolution of the mean density of each of the different components of the universe – baryonic matter, dark matter, radiation, neutrinos, dark energy, etc. (G is Newton's gravitational constant). The effect of spatial curvature parametrized by k: $k = 0$ indicates a flat universe.

The density of each component, ρ, evolves with the scale factor a as

$$\rho\,(a) \propto a^{-3(1+w)}\,, \tag{3}$$

with w the (constant) equation of state of each given component.

More conveniently, each of the different components of ρ can be written in terms of energy density parameters Ω defined as a fraction of the "critical energy density" ρ_c, the current energy density of a flat ($k = 0$) universe

$$\Omega \equiv \frac{\rho}{\rho_c} = \frac{8\pi G\rho}{3H^2}\,. \tag{4}$$

Non-relativistic matter has an equation state of $w_\mathrm{M} = 0$. From (3), its energy density Ω_M will be diluted as the universe expands as a^{-3} or by the volume. Ultrarelativistic matter, such as radiation and neutrinos, has $w_\mathrm{rad} = 1/3$. Its energy density Ω_rad is diluted more quickly by the expansion than matter as a^{-4}, decreasing faster than a simple volume expansion as radiation has momentum and therefore a wavelength, stretched by a factor of a. The final component Ω_DE, dark energy, must have a strong negative pressure to explain the observed cosmic acceleration and hence a negative (but not precisely known) w. Equation (2) is then written as (in a flat universe with $k = 0$)

$$H^2\,(a) = H_0^2\left[\Omega_\mathrm{M}a^{-3} + \Omega_\mathrm{rad}a^{-4} + \Omega_\mathrm{DE}a^{-3(1+w)}\right]\,, \tag{5}$$

where w is now the equation of state of only the dark energy component. Here w is assumed constant; for a non-constant w, the final term is replaced by

$$\Omega_\mathrm{DE}\exp\left(3\int_a^1 \frac{da'}{a'}\left[1 + w\left(a'\right)\right]\right)\,. \tag{6}$$

The expansion history of the universe is therefore a "competition" between these different components (Fig. 1). At early times, from around 3 s after the Big Bang until an age of 50,000 years (a cosmological redshift of ∼3,500), the universe was dominated by radiation. As the universe expanded and the radiation energy density dropped off as a^{-4}, the universe entered a matter-dominated era, where the gravitational attraction due to matter caused a period of deceleration. The energy density due to matter falls as a^{-3}, and at an age of about 9 billion years (a redshift of ∼0.45)

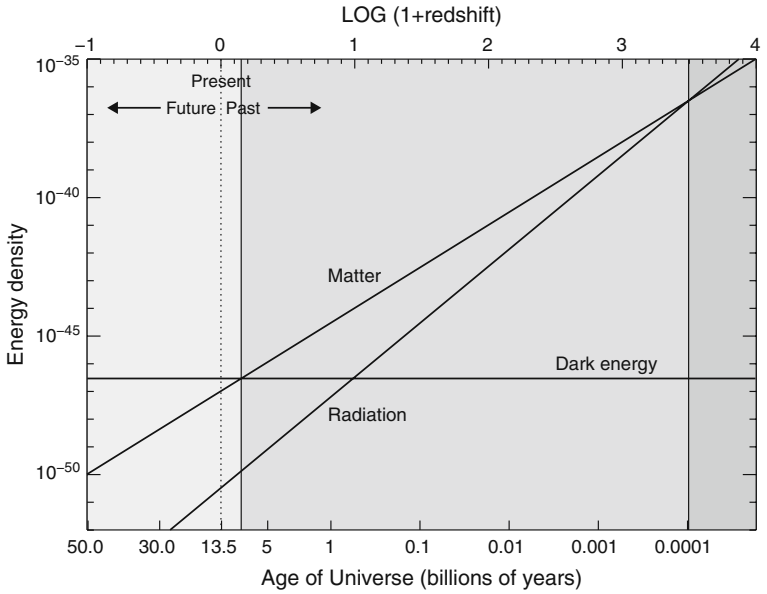

Fig. 1 The importance of the different components of the energy density of the universe as a function of the age of the universe and cosmological redshift for a flat universe. The three epochs discussed in the text are in different shades of *gray*. The early universe, at $z > 3{,}500$, is radiation dominated. Between $0.45 < z < 3{,}500$ is a matter-dominated era, and at $z < 0.45$ the universe is dark energy dominated

the effect of dark energy became dominant over that of gravity (although the effects of dark energy can be observed well before this redshift). Cosmic acceleration is, therefore, a relatively recent phenomenon in the expansion history of the universe, as dark energy was not important in terms of the expansion history at early times.

For the dark energy density term, the "simplest" solution is a cosmological constant, mathematically identical to a vacuum energy with a negative pressure exactly equal to its energy density unchanging with time: $w = -1$. This is equivalent to the Λ term introduced into GR by Einstein, and for that reason dark energy is often denoted by that term. In this case the expansion properties of a universe containing dark energy can be described by three parameters: w, $\Omega_{\rm DE}$, and $\Omega_{\rm M}$ (one parameter fewer if the universe is assumed flat and $\Omega_{\rm DE} + \Omega_{\rm M} = 1$). However, attempts to calculate the vacuum energy density from the zero-point energies of quantum fields result in estimates that are many orders of magnitude too large – a challenge to theories of fundamental physics.

Alternatively to vacuum energy, dark energy may be a scalar energy field of unknown physical origin which varies over both time and space, either decreasing or increasing in energy density, the latter leading to a "big rip" eventually tearing apart all structure. In these cases there is no a priori reason to assume that w is unchanging with redshift and many reasons to think that it is not.

The variety of possibilities to explain the cosmic acceleration make comparisons between observations and theory challenging. Ideally, the energy density of dark energy would be experimentally measured smoothly as a function of time, but in practical terms this is not yet possible. Instead, measuring $w(a)$ requires a parametrization of its form with a. The simplest is to assume w is constant: experiments then measure some average value $\langle w \rangle$. This is particularly valuable for assessing whether cosmic acceleration is consistent with vacuum energy (the cosmological constant): Is $\langle w \rangle$ consistent with -1? However, it is not particularly well motivated for other models of dark energy where w is expected to change with time.

For these "varying w" models, more complicated parametrizations should be used. Many simple, but useful, "two-parameter" parametrizations have been suggested with a linear dependence on either a (or redshift, z). The form $w(a) = w_0 + w_a(1 - a)$ is often used [15]. Other more general and complicated parametrizations are clearly possible [e.g., 16], including a principal component approach where $w(a)$ is measured over discrete intervals [17]. Other model-independent approaches such as direct reconstruction have also been examined [e.g., 18] and tested against real data [e.g., 19]. Each approach has advantages and drawbacks. Simpler parametrizations are easier to measure observationally, but harder to compare with models other than Λ. More complicated parametrizations clearly tend to result in noisier measurements.

As a final point, it should be noted that this framework is only relevant in the context of the FLRW metric and solutions to GR. If, instead, cosmic acceleration is an artifact or an indication of problems with GR, then this concept of w becomes meaningless. Typical approaches along these lines involve changes to the Friedmann equation (1) and (2) and the evolution of $a(t)$.

2.2 Distance Determination: The Standard Candle Method

Having established the simple framework necessary to measure dark energy, we now discuss how this can be achieved using distance estimation techniques. The principle underlying the use of standard candles to constrain the cosmological parameters through the observational effects of dark energy is to measure the expansion history $H(a) \equiv \dot{a}/a$ and compare with (5). The scale factor a is easy to measure via an object's redshift, z. When astronomical objects are observed, the photon wavelengths of the radiation that they emit are stretched ("redshifted") by the expansion of the universe by a factor $1/a = 1 + z$. The rate of change of a, $\dot{a}$ is more difficult (time is not a measurable!). Instead, distances to objects as a function of redshift are used, which are themselves intimately related to the expansion history. The comoving distance d (the distance between two points measured along a path defined at the present time) to an object at a cosmological redshift z is

$$d = \int_0^z \frac{c}{H(z')} \mathrm{d}z' = \frac{c}{H_0} \int_0^z \frac{\mathrm{d}z'}{\sqrt{\Omega_\mathrm{M} (1 + z')^3 + \Omega_k (1 + z')^2 + \Omega_\mathrm{DE} (1 + z')^{3(1+w)}}} , \tag{7}$$

where $H(z)$ is the Hubble parameter from (5), and the equation is written in terms of z rather than a. Related to this comoving distance are a variety of other distance definitions depending on the manner in which the distance measure is made. In particular, for SNe Ia the luminosity distance d_L is particularly useful:

$$d_L \equiv d(1+z)\,. \tag{8}$$

This luminosity distance can be *independently* estimated for an object of known intrinsic bolometric luminosity L (a standard candle) from an observation of the bolometric flux density f of the same object quite trivially as

$$d_L = \sqrt{\frac{L}{4\pi f}}\,. \tag{9}$$

The power of distance measures is clear: when L, f, and z are all known from measurements of a set of astrophysical objects, the only remaining unknowns in (7), (8), and (9) are the cosmological parameters. Thus, measuring a large set of astrophysical objects distributed in redshift which are known to be standard candles (such as SNe Ia) can directly measure parameters of interest and trace out the expansion history via the distance–redshift relation, traditionally on a Hubble diagram. In practice, even a knowledge of the absolute luminosity L is not required. Instead, *relative* distances between local and distant standard candles are sufficient, which has the considerable advantage of removing any dependence on H_0.

The size of the variation in the apparent magnitude[1] of a standard candle versus redshift for different cosmological models is shown in Fig. 2. For a simple measurement of $\langle w \rangle$, the "sweetspot" region is around $z = 0.6$ where the differences between various models are the largest, and the redshift is still small enough that high-quality observational data can be obtained. Above $z = 1$, the relative effect of a change in w in terms of apparent magnitude difference from that at $z = 1$ is very small: at these epochs the universe was decelerating and dark energy had only a minor influence on its evolution. Clearly, when trying to measure $w(a)$, samples of standard candles are required across the entire redshift range: the problem is quite degenerate if only a limited range in redshift can be observed. Figure 2 shows the variation assuming a simple linear function in $w(a)$.

A closely related technique to standard candles uses a different distance measure and the concept of "standard rulers," objects of known dimensions rather than known luminosity. Such sizes can be compared to the angular diameter distance d_A, the ratio of an object's (transverse) physical size to its angular size. It is related to the luminosity distance d_L as $d_A = d_L/(1+z)^2 = d/(1+z)$ and can probe the expansion history in a very similar way as standard candles. The method of measuring

[1] An astronomer's unit defined as $-2.5\log(f)$+constant. Note that smaller magnitudes represent brighter fluxes!

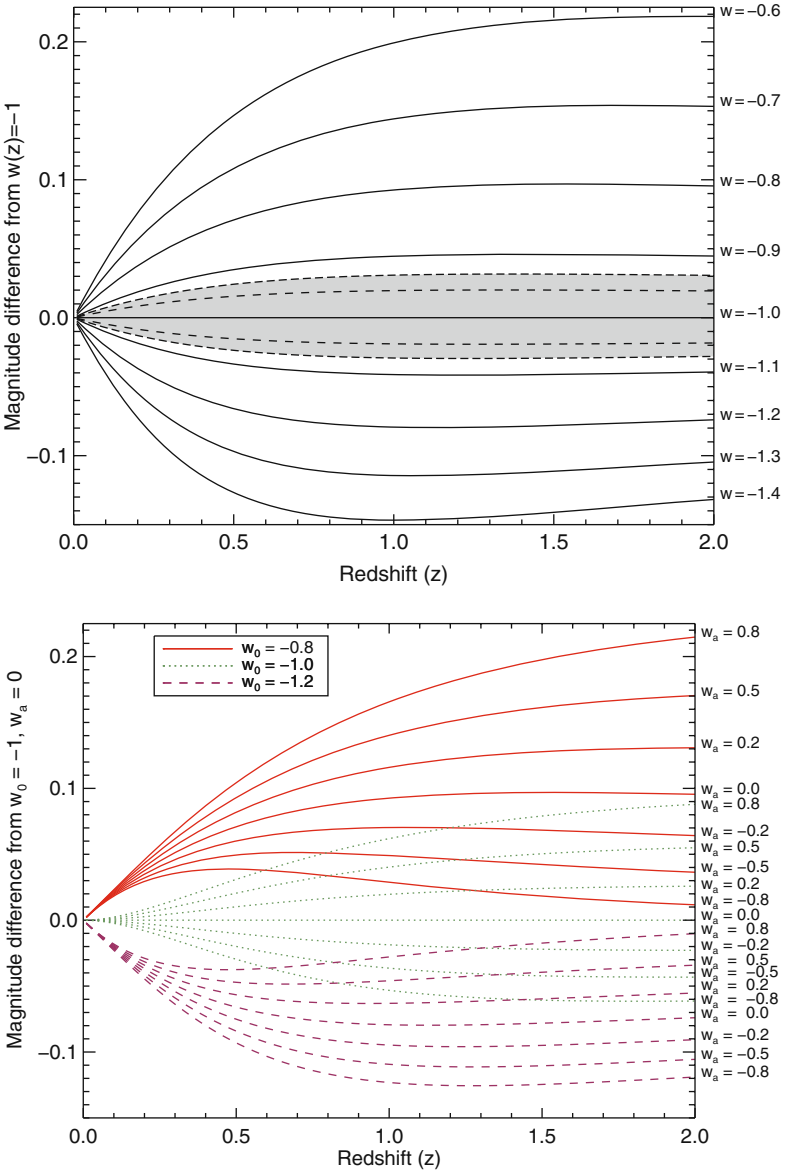

Fig. 2 The predicted variation in the apparent magnitude of a standard candle versus redshift for various cosmological models. In the *top frame* are different models assuming a constant equation of state, for $w = -0.6$ to $w = -1.4$. The current best constraints in $\langle w \rangle$ are shown in the *gray shaded* area. The upper *dashed–dot line* shows the constraints including systematic errors and the *dashed line* just the statistical error. The *bottom frame* is the same plot, but assuming a variable w according to $w(a) = w_0 + w_a(1 - a)$

baryon acoustic oscillations in the galaxy power spectrum, for example, exploits this idea (see Sect. 3.4). Generically, distance–redshift relations $d(z)$ provide very strong constraints on dark energy as they directly track the expansion history.

2.3 Type Ia Supernovae: A Brief Primer

One of the best standard candles known is Type Ia Supernovae, SNe Ia. Though an intimate knowledge of the physics of SNe Ia is not necessary for cosmological applications, it is extremely helpful in understanding the theoretical reasons as to why SNe Ia are such good standardized candles and in understanding the types of systematics that may affect the analysis.

The classification scheme for SNe – Ia, Ib, Ic, II, etc. – is mostly historical accident. Type I SNe were originally those events which did not show hydrogen in their spectra; Type II SNe have prominent H lines. The H-free Type I events were further sub-divided over time: Type Ia show strong Si absorption, Type Ib have no Si but instead He absorption, and Type Ic have no H, He, or Si. All SN types *except* SNe Ia are believed to be the result of the catastrophic core collapse of massive >8–$10\mathrm{M}_{\odot}$ stars.

SNe Ia have a different physical mechanism, the result of the thermonuclear destruction of a carbon–oxygen (C–O) white dwarf star residing in a binary system, i.e., with a nearby companion star. Having gained material from this companion via accretion, the mass of the white dwarf star becomes greater than that which can be supported by the electron degeneracy pressure of the white dwarf, at which point the core temperature of the white dwarf increases and C burning ensues in a sub-sonic deflagration. This C fusion raises the temperature further, an increase which cannot be regulated by the white dwarf star, where degeneracy pressure is independent of the temperature. The burning C fusion flame then accelerates into a supersonic detonation [20, 21] and a SN is the result.

Many details of the nature of the SN Ia explosion are still unclear, in particular the exact role of detonation versus deflagration. However, once the explosion has occurred, the resulting observed light curve – the luminosity evolution of the SN as a function of time – is broadly driven by relatively well-understood nuclear physics. The light curve is powered by the radioactive decay of ^{56}Ni, produced during the second or so of the SN explosion, into ^{56}Co, with a half-life of $\sim$6 days, and then by ^{56}Co into ^{56}Fe with a half-life of $\sim$77 days [e.g., 22], a process that can be confirmed observationally using line ratios in the late-time SN Ia spectra [23]. This radioactive decay deposits energetic gamma rays into the SN ejecta, which is heated and radiates thermally to produce the light curve that we observe. The mass of ^{56}Ni produced, M_{Ni}, is therefore the primary determinant of the peak brightness of the SN event. Observations indicate that for normal objects, M_{Ni} span the range 0.4–$0.9\mathrm{M}_{\odot}$, with a typical value of $0.6\mathrm{M}_{\odot}$. Photometrically, SNe Ia rise to maximum light in a period of approximately 20 days, followed by a rapid decline of about three magnitudes in the first month following maximum light and approximately one magnitude per month thereafter.

Though this consensus of an exploding near-Chandrasekhar mass C–O white dwarf star residing in a binary system is seldom seriously debated, the exact configuration of the progenitor system is more controversial. The companion star to the progenitor white dwarf could be a second white dwarf star ("double degenerate") or a main-sequence or giant star ("single degenerate"). Evidence from observations or theory for and against these two possibilities is ambiguous. These uncertainties are some of the biggest drawbacks of the SN Ia technique – in the absence of any theoretical guidance, it must be assumed that any variation can be empirically controlled (see Sect. 3).

Theoretical modeling of a SN Ia explosion is a complex problem, requiring the interior physics of the exploding white dwarf star to be related to what is finally observed: the light curves and spectra. This must be achieved by radiative transfer calculations, an unsolved problem in many astrophysical applications; SNe Ia are no exception. Unlike most astrophysical objects, SNe Ia contain no hydrogen, and therefore the opacities are always dominated by electron scattering (in the optical) or by a vast number of atomic lines (in the ultraviolet) making detailed predictions more challenging. The radiation transport in SNe Ia is non-local and the methods used in models of stellar atmospheres need refinement. The deposition of energy from the light curve and the explosion itself are also likely to be non-symmetrical. Given these complications, much of our understanding of SNe Ia is currently observationally motivated, though recent theoretical progress is now starting to provide theoretical insight [24].

2.4 Discovery of the Accelerating Universe

Type Ia Supernovae (SNe Ia) are apparently ideal as standard candles – they are bright, uniform, and possess a convenient approximately month-long duration during which they can be found and observed. This makes them extremely observationally attractive and practical as calibrateable standard candles, a realization that goes back many decades [25]. Yet, for many years following the realization of this potential, finding distant events in the numbers required for meaningful constraints was a considerable logistical and technological challenge. Years of searching were required to discover only a handful of distant SNe Ia [e.g., 26, 27]. The field only came of age through improving technology: the advent of large format CCD cameras on 4-m class telescopes, capable of efficiently scanning large areas of sky, and the simultaneous development of sophisticated image processing routines and powerful computers capable of rapidly analyzing the volume of data produced.

The substantial search effort culminated in the late 1990s when two independent surveys for distant SNe Ia [27, 28] made the same remarkable discovery: the high-redshift SNe Ia appeared about 40% fainter – or equivalently more distant – than expected in a flat, matter-dominated universe [29, 30, see Fig. 3]. This indicated that the expansion of the universe had been speeding up over the last $\sim$5–7 Gyr, providing compelling direct evidence for an accelerating universe. When these

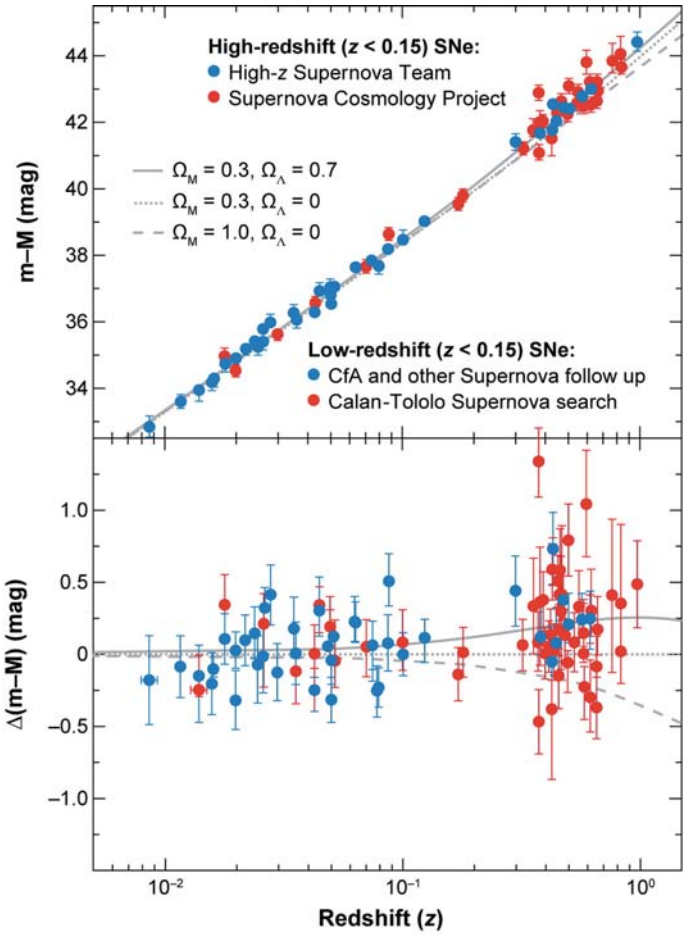

Fig. 3 The original "discovery data" which directly indicated the accelerating universe. This is the original Type Ia Supernovae Hubble diagram compiled from data taken by the Supernova Cosmology Project [30] and the High-z Supernova Search Team [29]. The *lower panel* shows the residuals in the distance modulus relative to an open universe. The SNe Ia lie above and are inconsistent with (fainter than) the non-accelerating universe lines. Reprinted, with permission, from the Annual Review of Astronomy and Astrophysics, Volume 46, © 2008 by Annual Reviews `www.annualreviews.org`

observations were combined with analyses of the cosmic microwave background, a consistent picture emerged of a spatially flat universe dominated by a "dark energy" responsible for ∼70–75% of its energy, opposing the slowing effect of gravity and accelerating the universe's rate of expansion.

Since these original observations, the number of cosmologically useful SNe has dramatically increased. The next section discusses how SNe are typically used in

a cosmological application and shows the latest results from these next-generation samples.

3 Cosmological Applications

Despite the claim that SNe Ia are standard candles, this is in fact a hugely oversimplified statement. SNe Ia possess approximately a factor of two variation in their "raw" peak brightnesses – limiting their usefulness in cosmological applications. The key astrophysical development in the cosmological use of SNe Ia was the realization that their luminosities could be further homogenized, or standardized, using simple empirical techniques and correlations between SN luminosity and other variables. As such, SNe Ia are *standardizable*, rather than standard, candles. There are two relationships commonly used to standardize SN Ia luminosities, discussed in the next section.

3.1 Standardization

The most well-known relationship is the classic light-curve width–luminosity relationship (WLR), the "Phillips relation" [31]. Raw SN Ia peak luminosities are strongly correlated with the width of the SN light curve – intrinsically brighter SNe Ia typically have wider (i.e., slower declining) light curves. Equivalently, SN Ia light curves can be described using a "stretch" parameter s [27], which stretches or contracts a template SN light curve to match an observed light curve: high-stretch SNe have wider and brighter light curves than average, and low-stretch SNe have narrower and fainter light curves.

The physical origin of this relationship is not universally agreed upon, but must be related to the mass of ^{56}Ni, M_{Ni}, synthesized in the SN Ia explosion. Clearly, if more ^{56}Ni is synthesized, the SN will have an increased peak luminosity, and therefore a higher temperature. As the SN ejecta remain optically thick for several months following the SN explosion, the width of the light curve is therefore related to the timescale for photons to escape from the ejecta by diffusion. Therefore, the WLR could be explained if brighter SNe Ia have higher effective opacities, κ, and hence a longer diffusion time, which scales as $k^{1/2}$, meaning the photons take longer to escape and a broader light curve is the result.

Several related physical mechanisms have been postulated to explain the WLR [for an excellent summary, see 32]. The first is a simple temperature effect [e.g., 33]: SNe Ia with a larger M_{Ni} have higher temperatures, leading to more radiation at shorter wavelengths, in the ultraviolet rather than the optical. As line opacity is substantially higher in the blue, κ should increase with an increasing M_{Ni}. The second mechanism is related to ionization state [e.g., 34, 35]: the diffusion time becomes shorter when ultraviolet photons are able to fluoresce to longer wavelengths and escape, via interactions with Fe-group elements. This fluorescence is more efficient in singly ionized species, which are more common in cooler (smaller M_{Ni}) SNe.

A third mechanism is simply related to the ejecta composition [36]: A larger $M_{\rm Ni}$ implies more Fe-group elements and therefore larger opacities.

More recent work [32] suggests that while present, these three effects are of secondary importance, and the WLR is instead driven primarily by the blanketing from Fe II/Co II lines which appear earlier in cooler SNe with smaller $M_{\rm Ni}$, and therefore produce a faster decline rate. Contemporary models are now capable of qualitatively reproducing the observational WLR [32, 24].

The second, and less well-understood, relationship is between the SN Ia luminosity and the SN color – intrinsically brighter SNe Ia typically have a bluer optical color than their fainter counterparts [e.g., 37]. Here, color is measured as the difference in magnitudes (or ratio of fluxes) in the B[2] and V bands. Naively, this is exactly the sense that would be expected from the dimming effect of dust – dust along the line of sight to distant SN events should both redden *and* dim their spectral energy distributions (SEDs). However, there is a major complicating factor. If the dust in distant galaxies which host SNe Ia is assumed to be the same as that present in the Milky Way, the result should be a relationship between SN color and luminosity much steeper than that observed. Multiple studies show that the Milky Way value for the ratio of total to selective extinction in the rest-frame B-band (R_B) of 4.1 is not consistent with that found when analyzing SNe Ia, with all evidence pointing to an effective $R_B < 4.1$: the color corrections on SN Ia are *incompatible with known galactic dust properties* [e.g., 39–41].

This suggests one of three possibilities. The first is that dust in galaxies which host SNe Ia is radically different to dust in the Milky Way. However, observations of quasars behind foreground galaxies can be used to probe extinction laws in other galaxies and show little evidence for this effect [42]. A second possibility is that the circumstellar dust surrounding the SN Ia progenitor white dwarf may play an important role [43]. Finally, there may be an additional intrinsic relationship between SN color and luminosity that does not correlate with light-curve shape. Interestingly, recent SN simulations do show evidence for intrinsic color–luminosity relationships in the same sense as that observed in the data [24].

So far, stretch and color are the only two photometric parameters that have been found to correlate with SN Ia luminosity, and nearly all cosmological analyses exploit these relationships, albeit in varying forms. The two relationships are shown in Figs. 4 and 5 using a sample of modern SN Ia data. These relationships can be applied to observed peak magnitudes m and take the form

$$m_{\rm corr} = m + \alpha\,(s-1) - \beta\mathcal{C}\ , \tag{10}$$

where the stretch–luminosity is parametrized by α, and the color–luminosity relationship is parametrized by β. Applying these, or similar, calibrating relationships

[2] The B-band is an optical "filter" centered at ∼450 nm and with a full width of ∼100 nm. The V-band is redder at ∼550 nm. See [38] for typical bandpasses assumed for this and other standard optical filters.

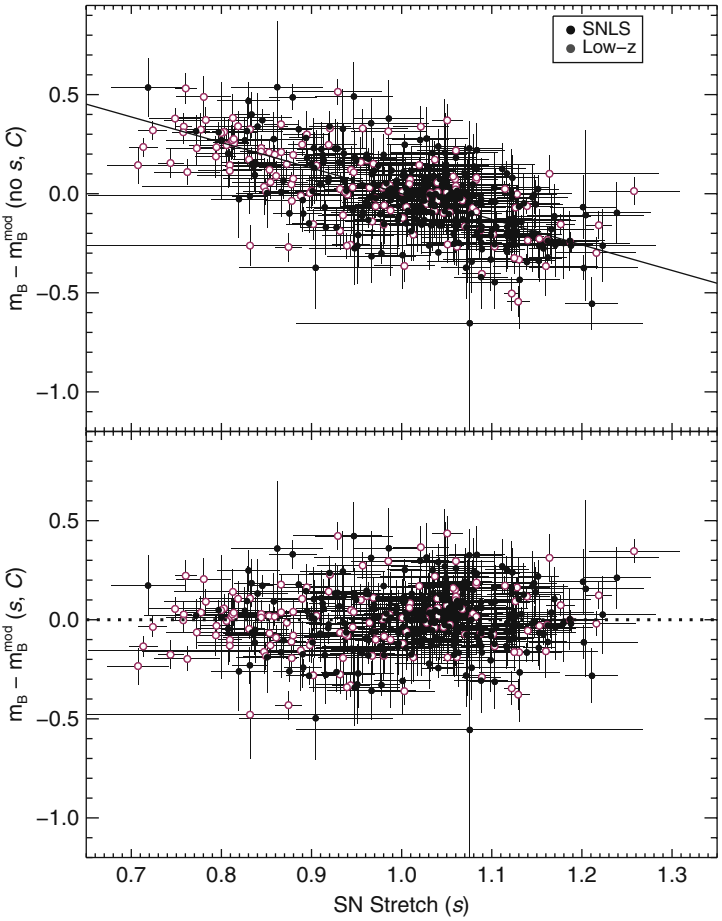

Fig. 4 The relationship between SN Ia luminosity and light-curve stretch. The *upper panel* shows the relationship, while the *lower panel* shows the trend between luminosity stretch after the best-fit relationship (*solid line* in *upper panel*) has been removed from the data. The *open circles* show data at low redshift, and the *filled circles* show data at high redshift

to SN Ia measurements provides distance estimates precise to ∼6–7% or 0.12–0.14 magnitudes, using various independent techniques [44–46].

3.2 Light-curve fitting

In nature, SNe Ia do not come neatly packaged in the form required for cosmological analyses. SNe are not usually observed exactly at maximum light, where they are best calibrated as standard candles, and the redshifting of their SEDs requires their observed fluxes be *k*-corrected back to the rest frame before they can be used in a cosmological context. Therefore, light-curve fitting is a critical component of a

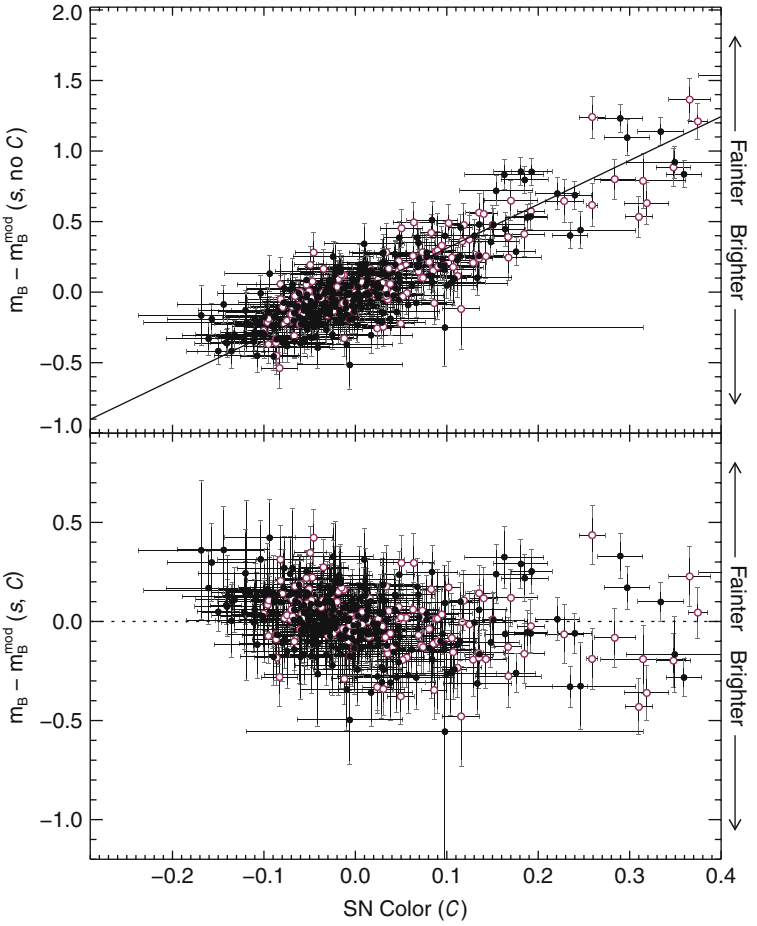

Fig. 5 As Fig. 4, but for SN C instead of stretch

cosmological analysis. This requires modeling of the SN light curves and spectra, and the importance of light curve fitting can be seen by considering the large number of methods described in the literature: a partial sample includes MLCS/MLCS2k2 [47, 44], stretch [27, 48, 49], Δm_{15} [50, 51], BATM [52], CMAGIC [53], SALT2 [45], and SiFTO [46]. These approaches vary considerably. For some SNe they produce very different results, but when a reasonably large sample is considered, the overall results are quite similar [54, 46].

The distinction between light-curve fitters and distance estimators is often not made explicit in the literature. The majority of the published packages are light-curve fitters, with the exception of MLCS/MLCS2k2 and BATM. A light-curve fitter attempts to find the best fit to a given set of observed SN Ia photometry. The parameters of this fit are then usually converted into a distance estimate, but technically this step is not part of the light-curve fit. A distance estimator attempts to find the distance directly rather than trying to obtain the best fit to the data. In both cases, only a relative distance is required.

The advantage of a distance estimator is that the output is what is desired for *most* applications of SN Ia data. Therefore, the products of such an analysis are simpler to use, and in principle such an approach may do a better job extracting the information directly relevant to this purpose. Their primary disadvantage is that, by their nature, distance estimators must use distance information in their training, usually in the form of residuals from the best-fit Hubble relation. This makes it difficult to include very low-redshift SNe which are not in the smooth Hubble flow and high-redshift data since here the residuals are strongly dependent on the cosmological model. To use such data properly, it is necessary to re-train the model from scratch for every value of the cosmological parameters one evaluates, which will be extremely computationally expensive. Therefore, in practice, light-curve fitters have access to a larger data set for training purposes. Neither approach is a priori obviously superior, though the advantages of incorporating data from a range of redshifts in the training does make light-curve fitters compelling. In particular, light-curve fitters are able to make use of the blue/UV part of the SN Ia SED by using high-redshift SN data, where the rest-frame UV is redshifted into optical bandpasses. Using this data in the model training can allow a drastic improvement in the color measurement for the most distant SN events. Depending on the data set, both light-curve fitters and distance estimators can be used with a typical precision of ∼6–10%.

The goal of both light-curve fitters and distance estimators is to reduce each SN Ia to a set of parameters that can then be used in cosmological fits. Here, we concentrate on light-curve fitters, though the same basic principles apply to distance estimators. A light-curve fitter such as SALT2 or SiFTO estimates, for each SN, its peak magnitude in some rest-frame passband, a measure of its light-curve shape and a measure of its color. Here we use the peak magnitude in the rest-frame B-band, m_B, the stretch s, and the $B-V$ color at the time of maximum light in the rest-frame B-band, $\mathcal{C}$. In principle, any rest-frame passband and color could be used; in practice the chosen passband should always correspond to a part of the SN Ia SED that is directly measured at all redshifts under study, which currently makes the B-band the most practical.

An integral part of any light-curve fitter (or distance estimator) is the k-correction. This converts a flux in an observed passband (e.g., r) into one in a rest-frame passband, in this case one common to all SNe (e.g., B). Given an SED model of the SN Ia ϕ_{SN}, parametrized by wavelength λ and epoch t, the rest-frame apparent magnitude m of a SN in filter B with a total system response T_B is given by

$$m_B = -2.5 \log_{10} \left[\frac{\int \phi_{\mathrm{SN}}(\lambda,t)\, T_B(\lambda)\, \lambda D\lambda}{\int \phi_{\mathrm{REF}}(\lambda)\, T_B(\lambda)\, \lambda D\lambda} \right] . \tag{11}$$

where ϕ_{REF} is the SED of some known and calibrated photometric reference SED, which places the measurement onto a known photometric system. This reference SED is classically chosen to be the A0V star, α-Lyrae or Vega.

In practice, experimentally Vega is a poor choice of flux reference standard – at an apparent magnitude of zero, it is far too bright to be observed by most telescopes, saturating CCD detectors. Calibration is, therefore, performed using observations of secondary stars that have previously been to a particular system, e.g., the Landolt

system [55], close to a Vega-based system. Unfortunately, Vega is quite a blue star and is unrepresentative of most of the secondary stars that have been chosen for observational calibration purposes. Modern precision calibration often, therefore, makes use of alternative flux reference SEDs to Vega, such as that of BD 17° 4708, which are fainter, have been calibrated to a higher precision, and have colors more similar to that of the secondary standards.

For the same SN observed in a passband r at a heliocentric redshift $z_{\rm hel}$, the observed apparent magnitude m_r is

$$m_r = -2.5 \log_{10} \left[\frac{1}{1+z_{\rm hel}} \frac{\int \phi_{SN}\,(\lambda,t)\, T_r\,(\lambda/\,(1+z_{\rm hel}))\,\lambda D\lambda}{\int \phi_{\rm REF}\,(\lambda)\, T_r\,(\lambda)\,\lambda D\lambda} \right] . \tag{12}$$

The k-correction is then defined as the difference between the two, namely

$$kc = m_r - m_B\ . \tag{13}$$

Thus, a conversion between observed and rest-frame magnitudes can be achieved.

These equations have two important implications. First, the SED of a typical SN Ia must be well known, over the entire light curve or phase for which there are photometric observations. SNe Ia demonstrate significant spectral evolution with time, and thus many observed SN Ia spectra are required to construct a spectral template suitable for light-curve fitting. Considerable work over the last decade has ensured that these templates are now generally available [56, 57]. The second implication is that a well-measured photometric reference SED must exist, for which magnitudes in a given passband must be known (see discussion above), allowing synthetic photometry to be performed for comparison to observed measures. Any uncertainty in which this SED is calibrated will directly impact as a systematic in the subsequent use of the peak magnitudes (see Sect. 4.1).

The light-curve fit is then a simple χ^2 minimization between the observations and synthetic magnitudes calculated according to (12), producing a peak magnitude, stretch, and color (together with their errors and covariances) which can be carried forward, for each event, to the cosmological parameter estimation stage, described in the next section.

3.3 Cosmological Parameter Estimation

A typical SN Ia cosmological fit minimizes an equation of the form

$$\chi^2 = \sum_N \frac{\left(m_B - m_B^{\rm mod}\,(z;\alpha,\beta,\mathcal{M}_B;\Omega_{\rm M},\langle w\rangle,...)\right)^2}{\sigma_{\rm stat}^2 + \sigma_{\rm int}^2}\ , \tag{14}$$

where we only include statistical errors for clarity. $\sigma_{\rm stat}$ is the total identified statistical error and includes uncertainties in both m_B and $m_B^{\rm mod}$, $\sigma_{\rm int}$ parametrizes the

intrinsic dispersion of each SN (see below), and the sum is over the N SNe Ia entering the fit. m_B are the maximum-light SN rest-frame B-band apparent magnitudes output from the light-curve fitter, and $m_B^{\rm mod}$ are the model B-band magnitudes for a SN at a redshift z given by

$$m_B^{\rm mod} = 5\log_{10}\mathcal{D}_L\left(z\right) - \alpha\left(s-1\right) + \beta\mathcal{C} + \mathcal{M}_B\ , \tag{15}$$

where $\mathcal{D}_L$ is the c/H_0 reduced luminosity distance d_L of (8). The c/H_0 factor in d_L is absorbed into $\mathcal{M}_B$; c is the speed of light. Explicitly, $\mathcal{M}_B = M_B + 5\log_{10}\left(c/H_0\right) + 25$, where M_B is the absolute magnitude of a SN Ia in the rest-frame B-band. Neither H_0 nor M_B need be assumed during the fitting process. The α and β nuisance parameters describe the stretch and color–luminosity relationships (Figs. 4 and 5) as described in (10). Any linear variation between SN color and stretch is also allowed for and is absorbed into the α term. The $\sigma_{\rm stat}$ term includes identified statistical errors affecting each SN. This typically comprises the statistical error in m_B from the light-curve fit, the statistical error in $m_B^{\rm mod}$ (essentially $\alpha\sigma_s$ and $\beta\sigma_{\mathcal{C}}$), a peculiar velocity error, the measurement error in each SN redshift projected into magnitude space, the uncertainty from Milky Way extinction corrections applied to m_B, and the covariances between s, $\mathcal{C}$, and m_B, which are correlated for an individual SN. $\sigma_{\rm stat}$ must be updated during the fitting process as α and β nuisance parameters are altered.

The $\sigma_{\rm int}$ term parametrizes the extra dispersion in m_B above and beyond the statistical and model uncertainties required to give a reduced χ^2 of one in the cosmological fits [e.g., 28]. This "intrinsic" dispersion could arise from unidentified sources of error in the analysis, but more likely is due to the imperfect nature of SNe Ia as standard candles – the α and β corrections do not completely eliminate the scatter in the SN Ia magnitudes. As $\sigma_{\rm int}$ may also include contributions from unidentified experimental errors, there is no a priori reason for its value to be the same from SN sample to SN sample.

The best-fitting cosmological parameters can then be found by forming a grid over the parameters of interest and computing the χ^2 of (14) at every point, converting into a probability via $P \propto \exp\left(-\frac{1}{2}\chi^2\right)$, with the proportionality set by normalizing over the grid. The "nuisance parameters" α, β, and $\mathcal{M}_B$ are marginalized over when generating confidence contours in the parameters of interest. Note that $\mathcal{M}_B$ can be marginalized analytically [e.g., 58], but this is not possible for α and β (doing so can bias the values found). These three parameters should not be assumed or fixed in the fit, as the uncertainties in their values need to be propagated.

3.4 Complementarity with Other Probes

Several of the cosmological parameters can, in principle, enter the χ^2 calculation of (14): the matter density $\Omega_{\rm M}$, the energy density of dark energy $\Omega_{\rm DE}$ and its equation of state w, and potentially the amount of curvature in the universe Ω_k. Other

complementary observations are, therefore, useful in conjunction with SNe Ia which place constraints, or priors, on Ω_M (e.g., observations of large-scale structure) or Ω_k (e.g., observations of the cosmic microwave background). At the time of writing, the techniques in common use and therefore most useful are observations of the cosmic microwave background and measurement of baryon acoustic oscillations. We briefly discuss these in turn.

3.4.1 The Cosmic Microwave Background

The cosmic microwave background (CMB) is a nearly isotropic background radiation discovered in the 1960s [59] with a near-perfect black body spectrum, peaking in the radio with a temperature of $\simeq 2.7$ K. The radiation originates from the early universe when the universe was much hotter and denser, and almost entirely ionized – photons and baryons were tightly coupled to each other, opaque to radiation. Some 370,000 years after the Big Bang at $z \sim 1{,}100$, the universe had expanded sufficiently and adiabatically cooled to a temperature near 3,000 K where electrons and protons are able to (re)combine to form neutral hydrogen ("The epoch of recombination"), decoupling the photons and baryons. The photons, free from the baryons, then propagate through the universe and appear as the CMB. As the universe has expanded by a factor of $\sim$1,100 since the epoch of recombination when the CMB was emitted, CMB photons appear considerably less energetic, redshifted into the microwave spectral region.

The CMB is extremely isotropic, but there are small temperature fluctuations of the order of 1/1000th of 1%. Before recombination, any initial density fluctuations, or perturbations, excite gravity-driven sound wave or acoustic oscillations in the relativistic-ionized plasma of the early universe. The matter and radiation are attracted, by gravity, into these regions of high density. A gravitational collapse then follows until photon pressure support becomes sufficient to halt the collapse, causing the overdensity to rebound because of the finite pressure of the gas, generating acoustic waves. These two effects compete to create oscillating density perturbations, driven by gravity and countered by photon pressure. At recombination as the photons are decoupled, those photons originating in overdense regions will appear hotter than average, while those from less-dense regions will appear colder. These small density fluctuations in the universe at that time are, therefore, imprinted directly onto the photons of the CMB, appearing to us as small temperature fluctuations or a temperature anisotropy.

These temperature differences can be "routinely" measured from the CMB power spectrum, the fluctuation in the CMB temperature (anisotropy) as a function of angular scale on the sky. This angular power spectrum of the CMB temperature anisotropy [60, 61, Fig. 6] series of peaks and troughs arises from the gravity-driven acoustic oscillations of the coupled photon–baryon fluid in this early universe. In particular, a strong peak is seen in the power spectrum on an angular scale corresponding to the sound horizon (r_s, the maximum distance sound waves can travel before recombination), where a perturbation crossed this horizon at exactly the time of recombination – the scale that was first feeling the causal effects of gravity at that

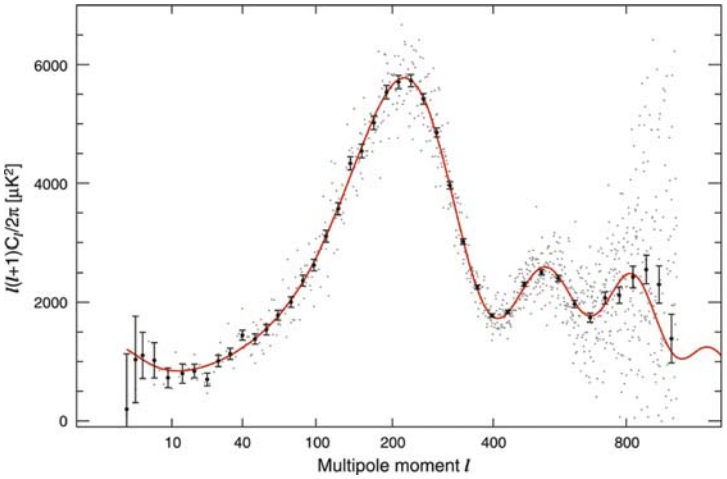

Fig. 6 The temperature anisotropy angular power spectrum from the WMAP-5 data [61]. The small gray dots represent the unbinned data, and the solid points the binned data with $1-\sigma$ error bars. The overplotted line is the best-fit ΛCDM cosmological model. From Dunkley et al. [61], Fig. 2. Reproduced by permission of the AAS

epoch. Smaller scales had been oscillating for longer and manifest as weaker peaks in the angular power spectrum.

A wealth of cosmological information is contained in positions and heights of the series of peaks and troughs [e.g., 62, 63]. The first peak, corresponding to the physical length of the sound horizon at recombination, depends on the curvature of space. If space is positively curved, then this sound horizon scale r_s will appear larger on the sky than in a flat universe; the opposite is true if space is negatively curved. The third peak can be used to help constrain $\Omega_{\rm M}$. However, the CMB by itself provides little direct constraint on dark energy – its role is a significant contribution in constraining Ω_k and $\Omega_{\rm M}$ (as well as the size of the sound horizon) for use in conjunction with other probes.

3.4.2 Baryon Acoustic Oscillations

Baryon acoustic oscillations (BAO) are closely related to the oscillations in the CMB angular power spectrum (Fig. 6). Following recombination, the immediate loss of photon pressure led to a consequent reduction in the effective sound speed of the baryons. The acoustic waves excited by the gravitationally unstable density fluctuations became "frozen" into the matter distribution with a characteristic size equal to their total propagation distance – the sound horizon scale r_s. This r_s can be seen in the power spectrum of the CMB temperature anisotropy, but additionally these sound waves remain "imprinted" in the baryon distribution and, through gravitational interactions, in the dark matter distribution as well. As galaxies (roughly) trace the dark matter distribution, observations of galaxy clustering can uncover this characteristic scale. Making this observation at different redshifts, therefore, allows

this scale r_s to be used as a standard ruler (Sect. 2.2) – just as SNe Ia trace $d(z)$ using $d_L(z)$, BAO measure $d_A(z)$ [e.g., 64, 65].

Power spectra analyses of galaxy redshift surveys contain the acoustic oscillations and are used to measure the cosmological parameters: the conversion of redshifts data into real space requires a cosmology to be assumed, and an incorrect choice will distort the power spectrum with the acoustic peaks appearing in the incorrect places. Observations of the CMB play a critical role here, as this same characteristic scale can be calibrated accurately by observations of anisotropy in the CMB imprinted at the same epoch. This observed, calibrated scale can, therefore, be used as a geometric probe of the expansion history – a measurement at low redshift provides an accurate measurement of the distance ratio between that redshift and $z \simeq 1{,}100$. Spectroscopic redshift BAO surveys can also measure the change of this characteristic scale radially along the line of sight as well as in the transverse direction, in effect a direct measurement of $H(z)$.

Measurements of the power spectra of galaxies are challenging. The oscillations appear as a series of bumps with an amplitude of only about 10%. This is substantially more subtle than the acoustic oscillations observed in the power spectrum of the CMB anisotropies because the impact of baryons on the far larger dark matter component is relatively small. Hence, enormous galaxy spectroscopic redshift surveys covering substantial volumes are required to make a constraining measurement. Photometric redshift surveys could, in principle, also be used and cheaply add hundreds of thousands of galaxies, this comes at the expense of a measurement of $H(z)$ and reduces the ability to measure $d_A(z)$ due to systematic errors and the higher noise of photometric redshifts over spectroscopic measures.

Although using BAOs to measure dark energy with precision requires enormous survey volumes and millions of galaxies, numerical simulations suggest that systematic uncertainties associated with BAO measurements are small – this method is currently believed to be (relatively) unaffected by systematic errors. The physics underlying the standard ruler can be understood from first principles. The main systematic uncertainties that are present in any interpretation of BAO measurements are the effects of nonlinear gravitational evolution and scale-dependent differences between the clustering of galaxies and of dark matter (known as bias). For spectroscopic redshift surveys, redshift distortions of the clustering can also shift the BAO features. However, studies suggest the resulting shift of the scale of the BAO peak in the galaxy power spectrum is 1% or less [e.g., 66].

3.5 Latest Cosmological Constraints

As might be expected, the quantity and quality of SN Ia data have dramatically improved since the original surveys. Dedicated allocations of observing time on 4-m class telescopes, such as the Canada–France–Hawaii Telescope (CFHT) and the Cerro Tololo Inter-American Observatory (CTIO) Blanco telescope, have provided homogeneous multi-color light curves of more than 500 distant SN Ia events over $z = 0.3-1.0$. The principle advances in this redshift range have come from the Supernova Legacy Survey [SNLS; 39] and ESSENCE [54]. At higher redshifts,

above $z = 1$, the *Hubble Space Telescope* has been used to locate ∼25 SN Ia events probing the expected epoch of deceleration [67, 68]. These latter observations also rule out the invocation of "gray dust" to explain the SN data in place of acceleration (see Sect. 4.1).

Lower redshift SN Ia samples are also important – and often neglected. The absolute luminosity of a SN Ia is not known precisely as the distance must be known independent of H_0 and so cannot be used a priori. The SN Ia method instead measures relative distances. Sets of local SNe at $0.015 < z < 0.10$, where the effect of varying the cosmological parameters is small, essentially anchor the analyses and allow relative distances to the more distant events to be measured. (At redshifts lower than $\simeq$ 0.015, the peculiar velocities of the SN Ia host galaxies, or bulk flows, can make the measurement both noisier and biased if not corrected for – see Sect. 4.1). The Sloan Digital Sky Survey (SDSS) SN survey is set to fill in the region from $0.1 < z < 0.3$, and many hundreds of lower redshift SNe Ia in the nearby Hubble flow ($0.03 < z < 0.1$) are either available or upcoming [e.g., 69–73]. The result of these new SN Ia data sets is a comprehensive set of well-calibrated events uniformly distributed from the local universe out to $z > 1$.

The current status of cosmological measurements of dark energy can be seen in Fig. 7 [74]. This uses around 500 SN Ia distributed in redshift from different SN Ia surveys, observations of the CMB from the WMAP-5 data release [61], and BAO measurements from the SDSS [75]. Using this combination of techniques, the latest results show that $\langle w \rangle$ is consistent with -1 with a sub-5% statistical precision. Systematics increase this total error to about 7% (see Sect. 4.2), though it should be cautioned that only systematics from the SN Ia analysis are included in this error estimate. Of particular note is that, at the present time, the BAO measurements provide an almost orthogonal constraint to SNe Ia in $\Omega_M/\langle w \rangle$ space (Fig. 7). Without

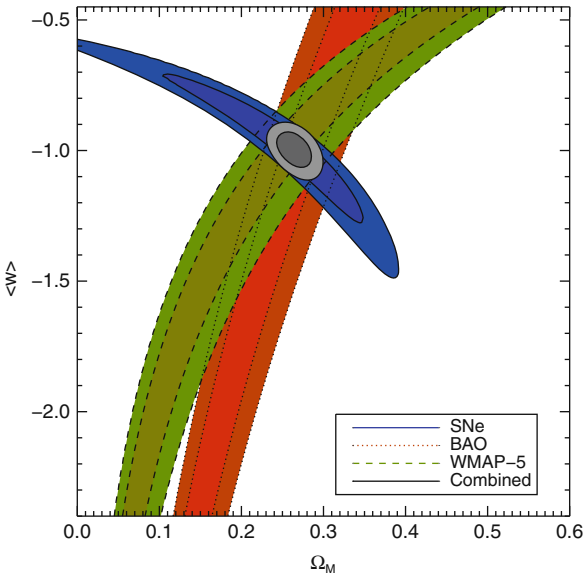

Fig. 7 Latest constraints on the nature of dark energy from SNe Ia and other techniques. The contours show the joint 1σ and 2σ constraints in $\langle w \rangle$ and Ω_M from SN Ia, baryon acoustic oscillations [75], and the cosmic microwave background WMAP-5 results [76]. A flat universe is assumed. The contours show statistical errors only and do not include systematic uncertainties

either the SN Ia constraints or the BAO, our measurement of the equation of state of dark energy would be considerably weakened. This is currently generically true – no single technique can yet place tight constraints on dark energy in isolation.

In the next section, we review the current systematic uncertainties which affect the SN Ia measurement and how these might be improved and reduced with upcoming surveys.

4 Systematics in SNe Ia

Though SNe Ia provided the first direct evidence for dark energy, and still provide the most mature and constraining measurements, there are a number of potential drawbacks to the technique – the apparently simple standard candle concept has several non-apparent difficulties. These difficulties are fundamentally related to the precision required (Fig. 2): detecting departures in dark energy from $w = -1$ requires an extremely sensitive experiment. A 10% difference in w from -1 is equivalent to a change in SN Ia brightness at $z = 0.6$ of only 0.04 magnitudes, or $<4\%$ in flux units, a level of precision perhaps not routinely achieved in astronomy. Therefore, systematic effects must be tightly controlled in any SN Ia experiment.

We discuss two broad classes of systematic error in this section. The first are identified systematics, those which are tractable, can be modeled, and the effects of which directly included in the cosmological fits. The second are the putative and more nebulous astrophysical systematics, related to uncertain configuration and physics of the SN Ia progenitor and explosion.

4.1 Identified Systematics

There are many different sources of identified systematic error in SN Ia experiments. The following is a brief summary of the most important at the time of writing. A full discussion can be found in [77].

Photometric calibration The required calibration of the SN Ia flux data onto a standard photometric system at a 1% accuracy is a challenging task [78] and remains the major source of systematic uncertainty in SN Ia experiments. This calibration must be controlled to the same level in both the high-redshift and low-redshift SN Ia samples. Calibration can be considered to consist of two steps: first, the observations must be standardized onto some photometric system by comparison to stars of known magnitude or flux [55]. Second, it is necessary to convert from this standard system into (relative) fluxes in order to compare SNe at different redshifts (and therefore observed at different parts of their SED). The first set is usually referred to as zero-point uncertainties the second as flux calibration uncertainties, which includes both bandpass uncertainties and uncertainties in the magnitudes of the chosen flux reference in those same bandpasses (see (11) and (12)).

Large, single instrument SN Ia surveys, such as those operating at high redshift, have a clear advantage here: they can devote considerably more effort to internal calibration than is practical for small surveys or those which make use of many instruments. In this regard, the zero-point uncertainty in calibrating to a standard system is usually quite small. Flux calibration uncertainties are more significant. In existing low-redshift surveys, the exact filter bandpasses are often not well measured. Further, calibrating low-redshift and high-redshift SN data onto the same system is a challenge as the observations are made in different filter systems (low-redshift data tend to be observed in the *UBVRI* photometric system, while high-redshift data are in the *griz* system). There are few standard stars with calibrated magnitudes in both systems, making this step one of the most uncertain in the analysis chain.

The color relation of SNe Ia Ideally, all of the measured colors for every SN Ia would be directly used in the distance estimate. In practice, this is difficult as the same rest-frame colors are not measured for all SNe even within the same survey. Therefore, most light-curve fitters make use of empirical relations between different wavelengths to build a single-color parameter for each SN ($\mathcal{C}$). These color relations have statistical uncertainties which are included in the final error budget. A particularly difficult issue relates to the fact that no SN color model perfectly reproduces the observed colors. That is, there is additional scatter in the relation between measurements at different wavelengths that is not explained by the measurement errors, and this intrinsic scatter must be accounted for in the color relations. The error on this scatter is very challenging to constrain and depends strongly on the functional form assumed.

SN Ia model uncertainties Even if all of the other systematics were absent there would still be some uncertainty in the SN Ia model used in the light-curve fitters as it is derived, or trained, from a limited set of data. This uncertainty takes a number of forms, such as errors in the SED (affecting the k-corrections), and errors in the relations that are used to combine different rest-frame colors into the color parameter, $\mathcal{C}$. Encouragingly, in implementations of the light-curve fitters SALT2 and SiFTO where the fitters use the same incidental settings (such as the filter response functions), the resulting fits on the same data are very similar [46]. Nonetheless, some differences do remain, and these should be included as a systematic error in the cosmological analysis.

Comparisons with distance estimators such as MLCS2k2 [44] are more problematic (see Sect. 3.2). Unlike SALT2 or SiFTO, MLCS2k2 explicitly attempts to separate intrinsic and extrinsic SN colors from photometric data, assuming that the extrinsic colors arise purely from dust, and that the remaining intrinsic color not related to the shape of the light curve does not affect the SN luminosity. SiFTO and SALT2 do not make this distinction. The merits of the two approaches depend critically on how well this separation can be performed and how well SN intrinsic color can be predicted by the light-curve shape. The former depends on accurate models of the distribution of extinction with redshift and how they combine with selection effects [54]. If these two conditions are met, then MLCS2k2, by incorpo-

rating additional information beyond SN photometry, may be able to give tighter statistical constraints on SN relative distances. A test of how well this procedure works is to check if the MLCS2k2 prediction of A_V after separation correlates with the residual from the Hubble diagram. Currently available versions fail this test [40], which indicates that, even if the second assumption is correct, the separation is not working correctly. While we fully expect that this situation will be addressed, especially in light of the improved low-redshift data sets that should soon be available, this means that comparing MLCS2k2 fits to SiFTO and SALT2 at the current time is not a useful way to study systematic errors.

Contamination by non-SN Ia Most SNe Ia used in cosmological analyses are spectroscopically typed, giving both a redshift and a firm indication that the object under study is a SN Ia and not some other transient event. Therefore, contamination by non-SNe Ia is expected to be minimal in most analyses. However, some is inevitable, particularly at the highest redshifts where the signal to noise of the confirmation spectra can be quite low. The principal contaminants are expected to be bright SNe Ib and Ic, the population demographics of which we know relatively little, which makes estimating the effects of any contamination difficult and imperfect. Reference [79] takes the approach of and modeling the population of bright SNe Ib and Ic as a Gaussian distribution in magnitude space with some width $\sigma_{\rm bc}$ and a mean offset of $\Delta_{\rm bc}$ from the SN Ia population. For current surveys, the effects on the cosmology are smaller than can be accurately measured, though any bias increases strongly with redshift where observations are larger and the fraction of SNe with more ambiguous spectra with CI 3 increases strongly. This may not be the case in future surveys that, due to their size, may have to rely on photometric rather than spectroscopic typing. Such efforts will require a much more precise understanding of the properties and demographics of SN Ib and Ic if they are to compete with spectroscopic surveys.

Malmquist bias Selection effects in SN Ia surveys, known collectively as "Malmquist bias," can also act as a systematic effect and must be included. At higher redshift, brighter (therefore bluer and higher stretch) SNe Ia will be preferentially discovered and followed spectroscopically, which would lead to a systematic brightening in the Hubble diagram residuals if left uncorrected. Fortunately, modern surveys such as SNLS can be simulated and the magnitude of this effect estimated. Corrections can then be applied directly to the SN Ia magnitudes and the uncertainties in these corrections included as a systematic error. Such simulations are substantially harder to perform on lower redshift surveys, which are not blind, rolling searches as at high-redshift, but instead tend to target known galaxies. This will be rectified in the next few years as a new generation of rolling low-redshift SN searches get underway.

Peculiar velocities The redshift lever arm needed to accurately measure the cosmic expansion requires the use of a local sample, and coherent large-scale local ($z < 0.1$) peculiar velocities add additional uncertainty to the Hubble diagram and hence to the derived cosmological parameters. It is possible to use local data to measure the local

velocity field and hence limit the impact on the derived cosmological parameters by "correcting" the measured redshifts of the local SN Ia host galaxies [80, 81]. Uncertainties in this correction are propagated through the cosmological fits as a systematic uncertainty.

Hubble bubble A related issue is the possibility of a monopole term in the local expansion — a so-called Hubble bubble. Recently, [44] found evidence for such an effect using light-curve fits to nearby SNe Ia. As discussed in [40], this is related to the interpretation of SN colors and the assumption that the relation between SN luminosity and color is well represented by a Milky Way dust law – the same issue is discussed in Sect. 3. The result is that when the same data is analyzed in a framework in which the relation between SN color and luminosity is determined from SN data, no evidence for a Hubble Bubble is seen.

Milky Way extinction correction SN Ia peak magnitudes are corrected for Milky Way extinction using the maps of [82]. In addition to the random error in their $E(B-V)$ values, there is a correlated uncertainty in the conversion from dust column density to extinction of 10%. This affects the nearby and distant SN differently because distant SN searches usually target regions of low galactic extinction and observe at longer wavelengths.

Gravitational lensing Gravitational lensing is expected to cause increased dispersion in the Hubble diagram of high-redshift SNe. While the mean amount of magnification is unity, when there are a small number of SN in each redshift bin, the asymmetric nature of the lensing probability, coupled with selection effects, can produce biases in the peak luminosities. For surveys over very small areas, lensing will also induce correlations between different SNe. These issues are studied in [83], who find that for current surveys, the number of SNe in each redshift range, and the survey area are large enough that these issues are minor. Lensing does induce additional, almost uncorrelated, scatter in the peak magnitudes, which can be included in the statistical error budget.

Gray dust Gray dust is a concept originally introduced to explain the faintness of the high-redshift SNe Ia when discovered in 1998, dust with negligible telltale reddening or additional dispersion. In its simplest form it can be easily tested against SN Ia observations, as it makes very different predictions for the Hubble diagram at $z > 1$. At these epochs, the SNe Ia probe the era of deceleration and will not get fainter at the same rate as that predicted by simple gray dust models [67].

A more pernicious kind of gray dust has been suggested by [84]. Theses are "replenishing dust" models, in which a constant density of gray dust is replenished at just the rate it is diluted by the expanding universe. Such models are virtually indistinguishable from that of an accelerating universe by just using the distance–redshift relation, as the dimming is directly proportional to distance traveled and mathematically similar to the effects of a cosmological constant. Dust of this sort with the required opacity, replenishing rate, and ejection velocity from galaxies ($>1{,}000\,\mathrm{km\,s^{-1}}$ for it to fill space uniformly without adding detectable dispersion) is virtually undetectable in the Hubble diagram of SNe Ia. However, such dust models

involve a large amount of fine tuning and appear contrived to be a simple alternative to dark energy.

4.2 Implementation of Systematic Errors

To include these identified systematic errors in the cosmological fits, (14) can be generalized by constructing a covariance matrix $\mathbf{V}$ to replace the σ terms. $\mathbf{V}$ is the combination of a systematics covariance matrix $\mathbf{V}_{\rm syst}$ and a (diagonal) statistical covariance matrix $\mathbf{V}_{\rm stat}$ generated from the statistical errors described above. We then minimize the χ^2 according to

$$\chi^2 = \sum_N \left(\mathbf{m_B} - \mathbf{m_B^{mod}}\right)^T \mathbf{V}^{-1} \left(\mathbf{m_B} - - - \mathbf{m_B^{mod}}\right) . \tag{16}$$

This (clearly) makes the cosmological fits more computationally expensive, but allows the uncertainties on the fit parameters to directly include systematic errors, as well as correctly accounting for systematics which induce correlations between different SNe and thus alter the position of the best-fit cosmological model.

The typical magnitude of the effect on the measurement of $\langle w \rangle$ of these identified systematic effects can be found in Table 1, and Fig. 8 shows the effect on the cosmological contours [77, 74]. The typical uncertainty in $\langle w \rangle$ increases from $\sim$4.5% when only considering statistical error to $\sim$7% when including systematics. Clearly, systematic errors are a large component of the total error budget – current constraints on $\langle w \rangle$ are systematics limited. However, there are many reasons to believe that this situation will radically improve as discussed in the next section.

Table 1 Example SN Ia systematic errors on $\langle w \rangle$ for a typical SN Ia sample

Systematic	Error in $\langle w \rangle$ (stat. + this sys.)[a]	Extra error from this systematic[b]
Stat. only	4.3	...
High-z zero points	4.5	1.3
Low-z zero points	4.7	1.9
Filter responses	4.5	0.9
Photometric reference colors	5.1	2.6
SN color relation	5.0	2.5
Peculiar velocities	4.4	0.5
Malmquist bias	4.4	0.7
Non-Ia contamination	4.4	0.7
Total	7.0	5.8

[a] The total error on $\langle w \rangle$ when considering the statistical only error, plus the additional error from this systematic.
[b] The additional error in $\langle w \rangle$ from this systematic error alone.

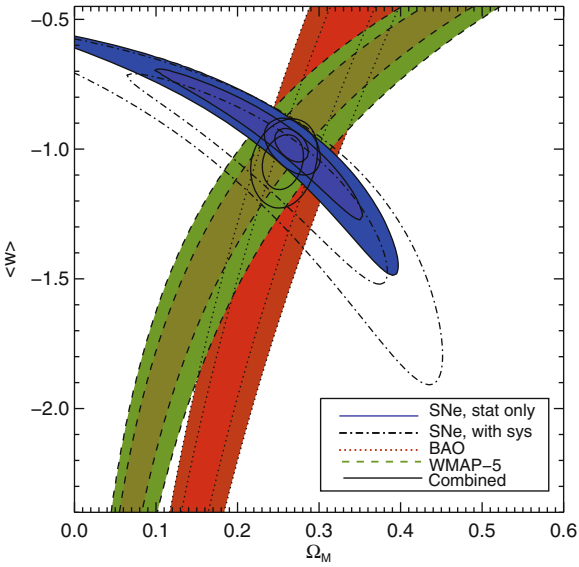

Fig. 8 As Fig. 7, but showing the effect of including SN Ia systematic errors when performing the cosmological fits. The *filled contours* are statistical only, and the *dashed contours* include the SN systematic errors

4.3 Prospects for Improvement

A careful reading of Sect. 4.1 will reveal that most of the dominant systematics are related to the current low-redshift data set and the necessity to cross-calibrate them to higher-redshift surveys such as SNLS. Since a large number of dramatically improved low-redshift data samples should become available in the next few years, observed in the same photometric system as the high-redshift data, it is worth asking exactly which aspects of these samples will help us reduce these errors.

The dominant systematics are the colors of the photometric reference standard BD 17° 4708, the differences between light-curve fitters and the zero points of the low-redshift samples. How will new low-redshift samples help with all of these? The uncertainties related to the color of BD 17° 4708 include large terms related to how well the Landolt magnitudes of this sub-dwarf can be transferred to the SNLS system. Because it is a slightly unusual star, and because the SNLS bandpasses are very unlike the Landolt ones used for measurements of the standard stars, these errors are large. Furthermore, the bandpasses of the Landolt system are simply not well understood, and never will be, putting a fundamental limit on how well this uncertainty can be characterized. Were the low-redshift sample replaced with one observed on a better understood system more similar to the SNLS or SDSS systems, these transfer effects would be minimized, substantially reducing the calibration errors. The low-redshift sample zero points can obviously be improved by better nearby samples. The effects on the SNLS/SDSS zero points are more subtle, but if these surveys are calibrated directly against a more similar system, they will be moderately reduced.

The differences between light-curve models will also be reduced with better training samples. While including high-redshift SN data in the training has been very useful, especially in the near-UV, new low-redshift samples can improve the situation even further. A particular lack in high-redshift data is multiple epochs of spectroscopy at different phases of the light curve. The uncertainty in the SN Ia color relations should also be substantially improved by new low-redshift samples simply because they dramatically increase the available pool of observations. These larger samples may allow an understanding of the various degeneracies between intrinsic and extrinsic SN Ia color by studying SNe Ia in a variety of environments, which could lead to qualitative improvements in SN Ia models.

The *Hubble Space Telescope* CALSPEC calibration used for BD 17° 4708 will also be refined and extended to more stars, particularly fainter examples that can be observed directly by modern survey telescopes, and of normal spectral types in the color range well sampled by standard star catalogs so that their magnitudes can be accurately transferred between systems. This program needs to be performed while the natural systems for high-redshift SN surveys (e.g., SDSS, SNLS) still exist; a major problem with the Landolt system is that its natural system does not – there will always be limitations as to how accurately any given flux standard can be tied to this system. This has implications for future absolute calibration programs. From the standpoint of SN observations, it is currently more important that the calibrated flux standard be tied to the magnitude system in use than how well the SED itself is calibrated, so good quality observations of fainter (and hence directly observable) standards are vastly preferable to improved observations of very bright standards such as Vega.

4.4 Further Astrophysical Systematics?

While the challenge of photometrically calibrating the physical SN Ia fluxes is considerable (Sect. 4.1), this is at least a well-defined and tractable problem on which substantial progress has been made. More concerning and pernicious is the possibility of intrinsic variability in the SN Ia population that cannot be empirically controlled. The most significant concerns are related to the unknown, or at least uncertain, astrophysical nature of the SN Ia events [e.g., 21].

Dramatic evolution in SN Ia properties is, however, ruled out. The spectra of SNe Ia are very similar across the entire redshift range in which they have been studied [e.g., 85–88], implying that the underlying physical process governing their explosions is not changing dramatically. Furthermore, SNe Ia in different type of host galaxies show very similar best-fitting cosmological models [89]. Different galaxy environments probe a large range of potential progenitor SN Ia stellar populations, from starburst galaxies with dominant young populations of stars, through normal galaxies with a substantial fraction of evolved stellar mass, to the old, evolved elliptical galaxies comprised of more homogeneous and old stellar populations. Hence their broad similarity among galaxy types is a powerful limit on the degree to which they can evolve.

However, it is well known that SNe Ia do have some connection with their host galaxy types. Although some SNe Ia occur in passive systems with no ongoing star formation, consistent with a long "delay time" from stellar birth to SN explosion, most events occur in star-forming galaxies, suggesting a shorter delay time [90, 91]. A strong correlation between the SN Ia rate and host galaxy star-formation rate is seen at all redshifts (Fig. 9). The most straightforward interpretation of this environmentally dependent SN Ia rate is a wide range of delay times, but the exact physical implications are unclear. For example, the SNLS relation between the SN Ia rate and star-formation rate implies that around 1% of all white dwarfs end their lives as SNe Ia [92], independent of their initial mass. As the single degenerate model typically has lower conversion efficiencies at lower masses, this suggests that some other mechanism is responsible for the production of at least some SNe Ia. However, the precise implication for the progenitor systems must await the construction of more detailed delay-time distributions, requiring more precise data on their host galaxies.

The "prompt" and "delayed" SNe Ia also possess different light curves. A trend of SN Ia luminosity – or equivalently stretch – versus galaxy morphological type has been observed [e.g., 93]: high-stretch SNe Ia are preferentially located in morphologically late-type galaxies (Fig. 10). This trend has also been observed when using host galaxy specific star-formation rates (sSFR; the SFR per unit stellar mass) instead of morphology [91]. The evidence suggests that prompt SNe Ia appear

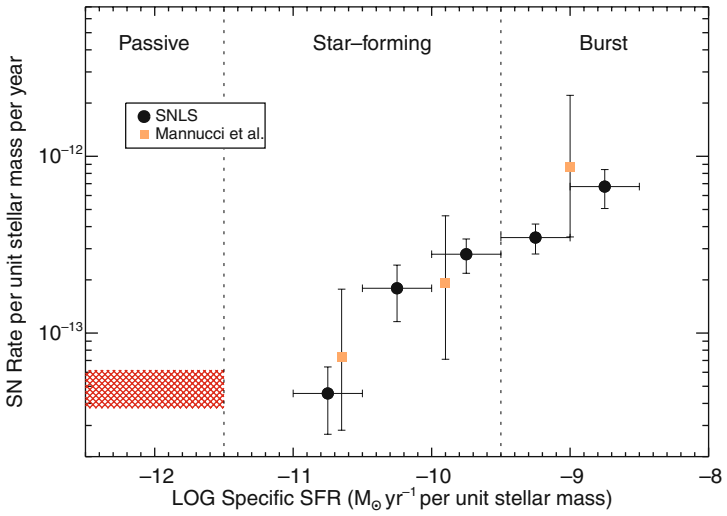

Fig. 9 The number of SNe Ia per unit stellar mass as a function of the star-formation rate (SFR) per unit stellar mass of the host galaxy. *Circles* represent SNLS data-points in star-forming galaxies. The *hashed area* shows the number per unit stellar mass as measured in the SNLS passive galaxies (assigned zero SFR). Shown for comparison (*square points*) is the evolution in SN Ia rate to later type galaxies observed locally by [90], normalized to the SNLS rate in passive galaxies. The *vertical dotted lines* show the division used to classify the host galaxies into different types. From Sullivan et al. [91], Fig. 6. Reproduced by permission of the AAS

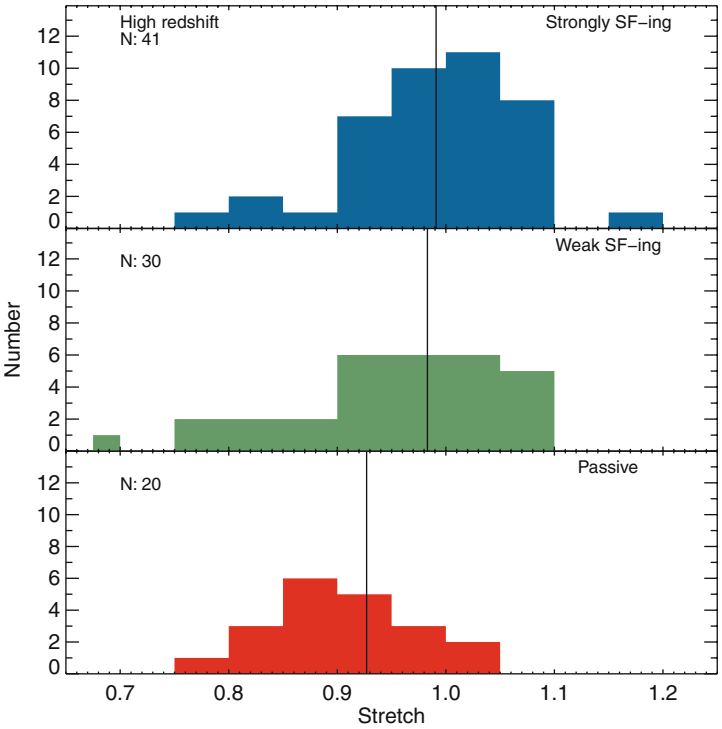

Fig. 10 The distribution of the SN Ia light-curve shape parameter "stretch" for SNLS high-redshift SNe, separated according to the specific star-formation rate of the host galaxy. The typical precision on the stretch measure is $\pm 0.01 - 0.02$, i.e., smaller than the bin width of the histograms. The *top panel* shows galaxies with a specific SFR (sSFR) of $\log(\text{sSFR}) > -9.5$, the *middle panel* galaxies with $-12.0 \leq \log(\text{sSFR}) \leq -9.5$, and the lower panel galaxies with $\log(\text{sSFR}) < -12.0$. The *vertical lines* show the positions of the median stretch in each histogram. From Sullivan et al. [91], Fig. 11. Reproduced by permission of the AAS

brighter with broader (higher stretch) light curves, while the delayed component are fainter with fast light curves – recall "stretch" is a key observable affecting the utility of SNe Ia as cosmological probes (see Sect. 3.1), correcting the luminosity of SNe Ia according to the width of their light curves. It is unclear whether the trends are primarily driven by progenitor age (passive systems hosting older stars) or progenitor metallicity (passive systems being the most massive and likely metal rich).

Regardless of the physical cause, the dependence of stretch on the apparent age of the SN Ia progenitor will lead to a subtle shift in the demographics of the SN Ia population. As the star-formation rate of the universe increases dramatically with look-back time, there will be many more prompt SNe Ia at high redshift than would be expected in the local universe. The consequence is that the high-redshift universe should have a more prevalent population of younger, higher stretch SNe than seen

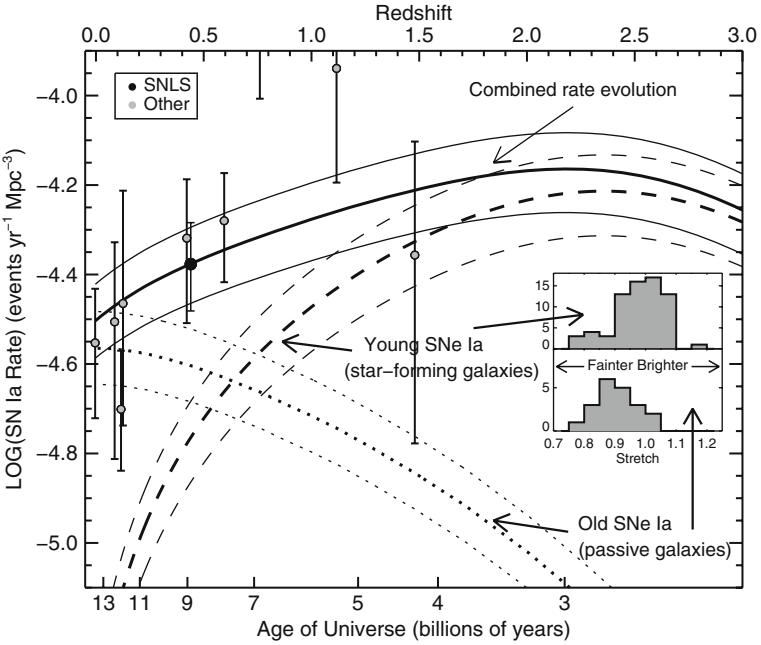

Fig. 11 The evolution of the volumetric SN Ia rate with redshift (*solid line*), based on a simple prompt (*dashed*) + delayed (*dotted*) SN Ia rate model [91]. The *inset* shows the distribution of the stretch of the SN Ia light curve associated with each of the two components — see also Fig. 10. This will lead to a subtle shift in the SN Ia stretch with redshift, as the demographics of the SN Ia population is changed

at low redshift – the mean stretch of the SN Ia population should evolve to be larger at higher redshift (Fig. 11).

Using data from various searches, this has now been tested using SN Ia data [94]. An increase in the average stretch with redshift is seen in the data, consistent with the empirical models, corresponding to $\sim$10% increase in the intrinsic SN Ia luminosity with redshift. Note that this difference is calibrated via the stretch–luminosity relationship, however, the degree to which the correction functions across the entire redshift range is still being studied.

There are also open questions as to how the metallicity or age of the progenitor star may influence the observed properties and luminosities of the SN Ia explosion, again leading to possible biases as the demographics of the SN Ia population shifts slightly with look-back time [94, 95]. Reference [96] argue dramatic changes in SN Ia physics with metallicity. They argue that synthesized ^{56}Ni mass should be linearly proportional to progenitor metallicity. Since the decay of ^{56}Ni drives the luminosity, SNe Ia in high metallicity environments should be less luminous. This is because stars from higher metallicity environments will end up with larger mass fractions of ^{22}Ne and ^{56}Fe after helium burning. Since these isotopes have excess neutrons, the authors argue that in these cases fewer radioactive elements are

produced during the process of burning to nuclear statistical equilibrium during a SN Ia. Note that these results are in sharp contrast to other studies which found no significant increase in ^{56}Ni with increasing metallicity [e.g., 97, 98]

Reassuringly, no definitive evidence for a dependence of SN Ia luminosity on inferred host galaxy metallicity has yet been uncovered [e.g., 99, 100]. Figure 12 shows the Hubble diagram residual as a function of inferred SN host galaxy metallicity for a sample of SNe Ia from the first year of the SNLS; no trend is apparent. Improved samples of SN data at both low and high redshifts should provide further constraints on the role of metallicity in influencing SN Ia luminosities.

An alternative approach is to explore the possible effects of progenitor metallicity through blanketing and wavelength-dependent features in the rest-frame UV spectra, corresponding to λλ 290–350 nm. This is a relatively unexplored region observationally as the atmosphere is opaque at these wavelengths: indeed the most complete studies have been conducted on high-redshift events where the UV spectral region is redshifted into the optical. Reference [97] argue that direct traces of the progenitor metallicity can best be seen in the unburned SN layers which are only observable significantly before maximum light. However, they also predict that an increase in progenitor metallicity will cause an increase in the amount of ^{54}Fe synthesized in the explosion, and this will result in an increase in line opacity in the UV region which may be observable at maximum light. The net effect is that an increased metallicity will result in an *increase* in the UV pseudo-

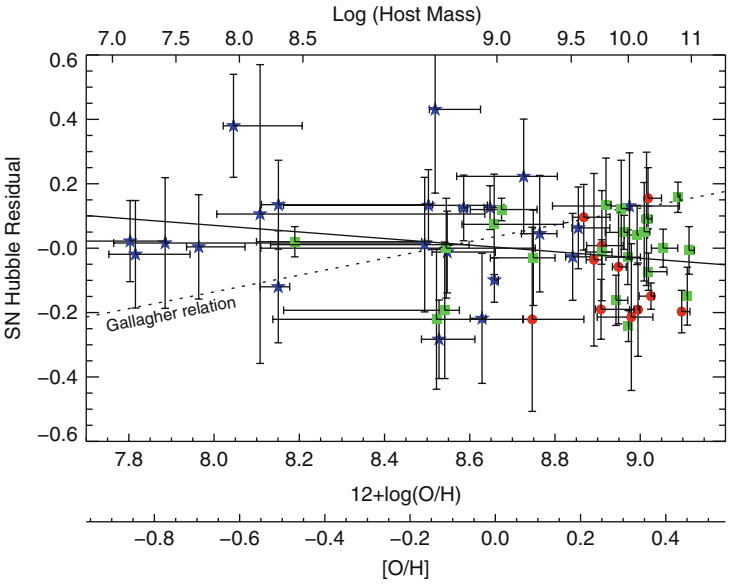

Fig. 12 Hubble residual versus inferred host galaxy metallicity for SNe Ia from the SNLS first-year sample [39]. The *solid line* is a fit to the data, consistent with zero at the 1σ level. The *dotted line* shows the relationship measured by [99], here ruled out by the data. From Howell et al. [100], Fig. 11. Reproduced by permission of the AAS

continuum at maximum light. Reference [101] examines the spectroscopic implications in greater detail. They simultaneously change the progenitor metallicity in the unburned C+O region and increase the amount of ^{54}Fe in the partially burned region. They find an increase in the level of the UV pseudo-continuum: as metallicity decreases so the line opacity decreases with the result that lines form deeper in the atmosphere and therefore from a lower velocity region. Such an effect is *opposite* to the effect predicted by [97]. However, [101] caution that the overall UV flux level is not necessarily a good indicator of metallicity, as it is dependent on many variables such as the temperature, density, and velocity of the C+O layer.

Various authors predict that the SN Ia rate should be affected by metallicity, although they do not make explicit predictions about the resulting effects on SN Ia properties. Reference [102] argues that in very low metallicity environments ($[Fe/H] < -1$) the white dwarf wind that they believe is essential for producing SNe Ia will be inhibited, thus leading to fewer SNe Ia. Reference [103] find that metallicity differences should alter the range of progenitor masses that produce SNe Ia.

In summary, therefore, theory cannot yet offer a clear consensus as to the effects of metallicity on SNe Ia. Indeed, there is disagreement not only about which effects are the most important, but also about the sign of any possible effect. This is a very challenging theoretical problem hindered by correlations between the wanted effect of metallicity and other correlations such as the viability of certain progenitor systems, the explosion mechanism and radiative transfer in an atmosphere under a variety of mixing conditions. Full simulations of all these effects may soon become feasible, but substantial campaigns will still be needed to track statistical shifts with metallicity.

Finally, there is the complex question of SN Ia colors. Their host galaxies span the full range of age from dwarf irregular galaxies through to giant ellipticals and contain vastly different amounts of dust. As discussed in Sect. 3.1, this dust has the effect of dimming the light from objects as it passes through, preferentially in the ultraviolet and blue spectral regions. But observed SN Ia properties seem inconsistent with known dust properties of the Milky Way. Probably, the question of dust represents the most serious challenge to SN Ia cosmology.

Observing SNe Ia at redder wavelengths where the effect of dust is smaller is one obvious potential solution. Observations of SNe in the rest-frame J and H bandpasses, which probe above 1 μm into the near infrared (IR), show a dramatically decreased dispersion on Hubble diagrams [e.g., 104]. SN Ia theory also suggests that any intrinsic variability in the population is smaller at these wavelengths [32]; in fact SNe Ia may be true standard candles without any need for a light-curve shape correction [24]. The problem with this approach is that near-IR observations are very challenging from the ground due to the increased sky brightness: to date observations are limited to a redshift of about 0.5. Space-based missions will be required to push this further for large samples of SNe.

While these potential systematics may appear serious, in part, this is because the SN Ia technique is the most mature and tested probe of dark energy. Despite many

decades of intensive searching, no fatal flaw has yet been identified – and SNe Ia have, so far, passed many detailed examinations of systematic effects with flying colors.

5 Concluding Remarks

Current SN Ia data sets play an integral part in making the most precise measurements of the dark energy driving the accelerating universe. At the same time, modern SN surveys also allow new insights into the astrophysics governing SN Ia progenitors. No effect has yet been uncovered that challenges the conclusions drawn from the use of SNe Ia in cosmological applications, but some open questions do remain. Why are the brightest SNe Ia associated with short delay times and young galaxies? How well do SNe Ia from different environments inter-calibrate in a cosmological analysis? A tantalizing possibility is the existence of *more than one progenitor mechanism*. The key to making progress is to pinpoint any fundamental environmental differences between the delayed and prompt events; several programs are underway to probe these questions.

In this regard, SN Ia observations seem in a particularly healthy state given the large number of proposed SN Ia surveys (Table 2). At low redshift, at least three new rolling SN Ia surveys are either planned or have commenced operation – the Palomar Transient Factory, the SkyMapper, and the La Silla SN Search. These surveys will generate many thousands of new SN Ia events, and ultimately studying these objects in details will lead to new insights into their progenitor systems and physics. At higher redshift, the Dark Energy Survey will eventually supplant the existing SNLS data set over the course of the next 5–7 years. The prospects for an eventual space-based mission to study dark energy are good, and SNe Ia are likely to be studied in detail by such an experiment.

As with any experimental technique, the final precision of the SN Ia cosmological results is governed by both statistical and systematic uncertainties. As discussed in this chapter, as more SNe Ia are used in the analysis and the statistical error decreases, the contribution of systematic errors has become increasingly important. Ultimately, the challenge of controlling systematics in SN cosmology is twofold.

Table 2 A selection of upcoming and planned SN Ia experiments

Time	$z < 0.1$	$0.1 < z < 1.0$	$z > 1$
Now	KAIT/CfA/CSP ∼300 SNFactory ∼100	SNLS ∼450 Essence ∼200 SDSS ∼ 250	HST/GOODS ∼20 HST/Clusters ∼20
2009–2013	SkyMapper PTF La Silla	DES PANSTARRS	HST/WFC3
2013++		LSST $\sim n \times 10^4$ JDEM	JDEM JWST TMT/ELTs

The first is photometric calibration; future, planned experiments will require a calibration of better than 1% in both the distant and nearby samples. The second is understanding the limitations of SNe Ia by investigating their astrophysical properties and, ultimately, controlling any subtle evolutionary effects that may emerge. This is the challenge for the next few years of SN Ia observation and theory.

Acknowledgments I acknowledge support from the Royal Society.

References

1. E. Hubble, Proc. Natl. Acad. Sci. **15**, 168 (1929). DOI 10.1073/pnas.15.3.168 60
2. E. Hubble and M.L. Humason, Astrophys. J. **74**, 43 (1931). DOI 10.1086/143323 60
3. P.J. Peebles and B. Ratra, Rev. Modern Phy. **75**, 559 (2003) 60
4. E.J. Copeland, M. Sami and S. Tsujikawa, Int. J. Modern Phys. D **15**, 1753 (2006). DOI 10.1142/S021827180600942X 60
5. A. Lue, R. Scoccimarro and G. Starkman, Phys. Rev. D **69**(4), 044005 (2004). DOI 10.1103/PhysRevD.69.044005 60
6. J. Peacock and P. Schneider, The Messenger **125**, 48 (2006) 60
7. A. Albrecht, G. Bernstein, R. Cahn, W.L. Freedman, J. Hewitt, W. Hu, J. Huth, M. Kamionkowski, E.W. Kolb, L. Knox, J.C. Mather, S. Staggs and N.B. Suntzeff, ArXiv Astrophysics e-prints (2006) 60
8. R. Trotta and R. Bower, Astron. Geophys. **47**(4), 040000 (2006). DOI 10.1111/j.1468-4004.2006.47420.x 60
9. J.A. Frieman, M.S. Turner and D. Huterer, Annu. Rev. Astron. Astrophys. **46**, 385 (2008). DOI 10.1146/annurev.astro.46.060407.145243 60
10. S.M. Carroll, W.H. Press and E.L. Turner, Annu. Rev. Astron. Astrophys. **30**, 499 (1992) 61
11. G. Efstathiou, W.J. Sutherland and S.J. Maddox, Nature **348**, 705 (1990). DOI 10.1038/348705a0 62
12. L.M. Krauss and M.S. Turner, Gen. Rel. Grav. **27**, 1137 (1995). DOI 10.1007/BF02108229 62
13. J.P. Ostriker and P.J. Steinhardt, Nature **377**, 600 (1995). DOI 10.1038/377600a0 62
14. W.L. Freedman, et al., Astrophys. J. **553**, 47 (2001). DOI 10.1086/320638 63
15. E.V. Linder, Phys. Rev. Lett. **90**(9), 091301 (2003). DOI 10.1103/PhysRevLett.90.091301 65
16. P.S. Corasaniti and E.J. Copeland, Phys. Rev. D **67**(6), 063521 (2003). DOI 10.1103/PhysRevD.67.063521 65
17. D. Huterer and G. Starkman, Phys. Rev. Lett. **90**(3), 031301 (2003). DOI 10.1103/PhysRevLett.90.031301 65
18. V. Sahni and A. Starobinsky, Int. J. Modern Phys. D **15**, 2105 (2006). DOI 10.1142/S0218271806009704 65
19. R.A. Daly, S.G. Djorgovski, K.A. Freeman, M.P. Mory, C.P. O'Dea, P. Kharb and S. Baum, Astrophys. J. **677**, 1 (2008). DOI 10.1086/528837 65
20. A.M. Khokhlov, Astron. Astrophys. **245**, 114 (1991) 68
21. W. Hillebrandt and J.C. Niemeyer, Annu. Rev. Astron. Astrophys. **38**, 191 (2000). DOI 10.1146/annurev.astro.38.1.191 68, 88
22. S.A. Colgate and C. McKee, Astrophys. J. **157**, 623 (1969). DOI 10.1086/150102 68
23. M.J. Kuchner, R.P. Kirshner, P.A. Pinto and B. Leibundgut, Astrophys. J. **426**, L89+ (1994). DOI 10.1086/187347 68
24. D. Kasen, F. Roepke and S.E. Woosley, ArXiv e-prints (2009) 69, 72, 93
25. C.T. Kowal, Astron. J. **73**, 1021 (1968) 69
26. H.U. Norgaard-Nielsen, L. Hansen, H.E. Jorgensen, A. Aragon Salamanca and R.S. Ellis, Nature **339**, 523 (1989). DOI 10.1038/339523a0 69

27. S. Perlmutter, et al., Astrophys. J. **483**, 565 (1997). DOI 10.1086/304265 69, 71, 74
28. B.P. Schmidt, et al., Astrophys. J. **507**, 46 (1998). DOI 10.1086/306308 69, 77
29. A.G. Riess, et al., AJ **116**, 1009 (1998). DOI 10.1086/300499 69, 70
30. S. Perlmutter, et al., The Supernova Cosmology Project, Astrophys. J. **517**, 565 (1999). DOI 10.1086/307221 69, 70
31. M.M. Phillips, Astrophys. J. **413**, L105 (1993). DOI 10.1086/186970 71
32. D. Kasen and S.E. Woosley, Astrophys. J. **656**, 661 (2007). DOI 10.1086/510375 71, 72, 93
33. P. Hoeflich, A. Khokhlov, J.C. Wheeler, M.M. Phillips, N.B. Suntzeff and M. Hamuy, Astrophys. J. **472**, L81+ (1996). DOI 10.1086/310363 71
34. P.A. Pinto and R.G. Eastman, Astrophys. J. **530**, 744 (2000). DOI 10.1086/308376 71
35. P.A. Pinto and R.G. Eastman, Astrophys. J. **530**, 757 (2000). DOI 10.1086/308380 71
36. P.A. Mazzali, K. Nomoto, E. Cappellaro, T. Nakamura, H. Umeda and K. Iwamoto, Astrophys. J. **547**, 988 (2001). DOI 10.1086/318428 72
37. R. Tripp, Astron. Astrophys. **331**, 815 (1998) 72
38. M.S. Bessell, PASP **102**, 1181 (1990). DOI 10.1086/132749 72
39. P. Astier, et al., Astron. Astrophys. **447**, 31 (2006). DOI 10.1051/0004-6361:20054185 72, 80, 92
40. A. Conley, R.G. Carlberg, J. Guy, D.A. Howell, S. Jha, A.G. Riess and M. Sullivan, Astrophys. J. **664**, L13 (2007). DOI 10.1086/520625 72, 84, 85
41. S. Nobili and A. Goobar, Astron. Astrophys. **487**, 19 (2008). DOI 10.1051/0004-6361:20079292 72
42. L. Østman, A. Goobar and E. Mörtsell, Astron. Astrophys. **485**, 403 (2008). DOI 10.1051/0004-6361:20079187 72
43. A. Goobar, Astrophys. J. **686**, L103 (2008). DOI 10.1086/593060 72
44. S. Jha, A.G. Riess and R.P. Kirshner, Astrophys. J. **659**, 122 (2007). DOI 10.1086/512054 73, 74, 83, 85
45. J. Guy, et al., Astron. Astrophys. **466**, 11 (2007). DOI 10.1051/0004-6361:20066930 73, 74
46. A. Conley, et al., Astrophys. J. **681**, 482 (2008). DOI 10.1086/588518 73, 74, 83
47. A.G. Riess, W.H. Press and R.P. Kirshner, Astrophys. J. **473**, 88 (1996). DOI 10.1086/178129 74
48. G. Goldhaber, et al., Astrophys. J. **558**, 359 (2001). DOI 10.1086/322460 74
49. R.A. Knop, et al., Astrophys. J. **598**, 102 (2003) 74
50. M. Hamuy, M.M. Phillips, N.B. Suntzeff, R.A. Schommer, J. Maza, R.C. Smith, P. Lira and R. Aviles, Astron. J. **112**, 2438 (1996). DOI 10.1086/118193 74
51. J.L. Prieto, A. Rest and N.B. Suntzeff, Astrophys. J. **647**, 501 (2006). DOI 10.1086/504307 74
52. J.L. Tonry, et al., Astrophys. J. **594**, 1 (2003) 74
53. L. Wang, G. Goldhaber, G. Aldering and S. Perlmutter, Astrophys. J. **590**, 944 (2003). DOI 10.1086/375020 74
54. W.M. Wood-Vasey, et al., Astrophys. J. **666**, 694 (2007). DOI 10.1086/518642 74, 80, 83
55. A.U. Landolt, Astron. J. **104**, 340 (1992). DOI 10.1086/116242 76, 82
56. P. Nugent, A. Kim and S. Perlmutter, Publ. Astron. Soc. Pacific **114**, 803 (2002). DOI 10.1086/341707 76
57. E.Y. Hsiao, A. Conley, D.A. Howell, M. Sullivan, C.J. Pritchet, R.G. Carlberg, P.E. Nugent and M.M. Phillips, Astrophys. J. **663**, 1187 (2007). DOI 10.1086/518232 76
58. M. Goliath, R. Amanullah, P. Astier, A. Goobar and R. Pain, Astron. Astrophys. **380**, 6 (2001). DOI 10.1051/0004-6361:20011398 77
59. A.A. Penzias and R.W. Wilson, Astrophys. J. **142**, 419 (1965). DOI 10.1086/148307 78
60. M.R. Nolta, et al., Astrophys. J. **180**, 296 (2009). DOI 10.1088/0067-0049/180/2/296 78
61. J. Dunkley, et al., Astrophys. J. **180**, 306 (2009). DOI 10.1088/0067-0049/180/2/306 78, 79, 81
62. P.J.E. Peebles and J.T. Yu, Astrophys. J. **162**, 815 (1970). DOI 10.1086/150713 79
63. J.R. Bond and G. Efstathiou, MNRAS **226**, 655 (1987) 79
64. C. Blake and K. Glazebrook, Astrophys. J. **594**, 665 (2003). DOI 10.1086/376983 80
65. H.J. Seo and D.J. Eisenstein, Astrophys. J. **598**, 720 (2003). DOI 10.1086/379122 80
66. H.J. Seo and D.J. Eisenstein, Astrophys. J. **665**, 14 (2007). DOI 10.1086/519549 80
67. A.G. Riess, et al., Astrophys. J. **607**, 665 (2004) 81, 85
68. A.G. Riess,et al., Astrophys. J. **659**, 98 (2007). DOI 10.1086/510378 81

69. M. Hamuy, et al., AJ **112**, 2408 (1996). DOI 10.1086/118192 81
70. S. Jha, et al., AJ **131**, 527 (2006). DOI 10.1086/497989 81
71. M. Hamuy, et al., PASP **118**, 2 (2006). DOI 10.1086/500228 81
72. M. Hicken, et al., Astrophys. J. **700**, 331 (2009). DOI 10.1088/0004-637X/700/1/331 81
73. A. Rau, et al., ArXiv e-prints (2009) 81
74. M. Sullivan, et al., in *Astrophys. J.,* submitted (2010) 81, 86
75. D.J. Eisenstein, et al., Astrophys. J. **633**, 560 (2005). DOI 10.1086/466512 81
76. E. Komatsu, et al., Astrophys. J. **180**, 330 (2009). DOI 10.1088/0067-0049/180/2/330 81
77. A.J. Conley, et al., Astrophys. J., submitted (2010) 82, 86
78. N. Regnault, et al., Astron. Astrophys. **506**, 999 (2009) 82
79. N.L. Homeier, Astrophys. J. **620**, 12 (2005). DOI 10.1086/427060 84
80. L. Hui and P.B. Greene, Phys. Rev. D **73**(12), 123526 (2006). DOI 10.1103/PhysRevD.73.123526 85
81. J.D. Neill, M.J. Hudson and A. Conley, Astrophys. J. **661**, L123 (2007). DOI 10.1086/518808 85
82. D.J. Schlegel, D.P. Finkbeiner and M. Davis, Astrophys. J. **500**, 525 (1998). DOI 10.1086/305772 85
83. D.E. Holz and E.V. Linder, Astrophys. J. **631**, 678 (2005). DOI 10.1086/432085 85
84. A. Goobar, L. Bergström and E. Mörtsell, Astron. Astrophys. **384**, 1 (2002). DOI 10.1051/0004-6361:20020002 85
85. I.M. Hook, et al., Astron. J. **130**, 2788 (2005). DOI 10.1086/497635 88
86. S. Blondin, et al., Astron. J. **131**, 1648 (2006). DOI 10.1086/498724 88
87. R.S. Ellis, et al., Astrophys. J. **674**, 51 (2008). DOI 10.1086/524981 88
88. M. Sullivan, R.S. Ellis, D.A. Howell, A. Riess, P.E. Nugent and A. Gal-Yam, Astrophys. J. **693**, L76 (2009). DOI 10.1088/0004-637X/693/2/L76 88
89. M. Sullivan, et al., MNRAS **340**, 1057 (2003) 88
90. F. Mannucci, M. della Valle, N. Panagia, E. Cappellaro, G. Cresci, R. Maiolino, A. Petrosian and M. Turatto, Astron. Astrophys. **433**, 807 (2005). DOI 10.1051/0004-6361:20041411 89
91. M. Sullivan, et al., Astrophys. J. **648**, 868 (2006). DOI 10.1086/506137 89, 90, 91
92. C.J. Pritchet, D.A. Howell and M. Sullivan, Astrophys. J. **683**, L25 (2008). DOI 10.1086/591314 89
93. M. Hamuy, S.C. Trager, P.A. Pinto, M.M. Phillips, R.A. Schommer, V. Ivanov and N.B. Suntzeff, Astron. J. **120**, 1479 (2000). DOI 10.1086/301527 89
94. D.A. Howell, M. Sullivan, A. Conley and R. Carlberg, Astrophys. J. **667**, L37 (2007). DOI 10.1086/522030 91
95. D. Sarkar, A. Amblard, A. Cooray and D.E. Holz, Astrophys. J. **684**, L13 (2008). DOI 10.1086/592019 91
96. F.X. Timmes, E.F. Brown and J.W. Truran, Astrophys. J. **590**, L83 (2003). DOI 10.1086/376721 91
97. P. Höflich, J.C. Wheeler and F.K. Thielemann, Astrophys. J. **495**, 617 (1998) 92, 93
98. K. Iwamoto, F. Brachwitz, K. Nomoto, N. Kishimoto, H. Umeda, W.R. Hix and F.K. Thielemann, Astrophys. J. **125**, 439 (1999). DOI 10.1086/313278 92
99. J.S. Gallagher, P.M. Garnavich, N. Caldwell, R.P. Kirshner, S.W. Jha, W. Li, M. Ganeshalingam and A.V. Filippenko, Astrophys. J. **685**, 752 (2008). DOI 10.1086/590659 92
100. D.A. Howell, et al., Astrophys. J. **691**, 661 (2009). DOI 10.1088/0004-637X/691/1/661 92
101. E.J. Lentz, E. Baron, D. Branch, P.H. Hauschildt and P.E. Nugent, Astrophys. J. **530**, 966 (2000) 93
102. C. Kobayashi, T. Tsujimoto, K. Nomoto, I. Hachisu and M. Kato, Astrophys. J. **503**, L155+ (1998). DOI 10.1086/311556 93
103. N. Langer, A. Deutschmann, S. Wellstein and P. Höflich, Astron. Astrophys. **362**, 1046 (2000) 93
104. W.M. Wood-Vasey, et al., Astrophys. J. **689**, 377 (2008). DOI 10.1086/592374 93

Modified Gravity Models of Dark Energy

S. Tsujikawa

Summary We review the recent progress for modified gravity models of dark energy – including $f(R)$ gravity, scalar-tensor theories, and braneworld models.
In $f(R)$ gravity, where the Lagrangian density f is a function of the Ricci scalar R, the coupling strength between dark energy and non-relativistic matter is of order 1 ($Q = -1/\sqrt{6}$) in the Einstein frame. Even in this situation it is possible for $f(R)$ models to be consistent with local gravity constraints under the chameleon mechanism, while at the same time satisfying conditions for the cosmological viability. We present a number of viable $f(R)$ models that satisfy cosmological and local gravity constraints.

We also study a class of scalar-tensor dark energy models based on Brans–Dicke theory with a scalar-field potential. The action in the Einstein frame can be viewed as a coupled quintessence scenario with a constant coupling Q that is related to a Brans–Dicke parameter $\omega_{\rm BD}$ via $3+2\omega_{\rm BD} = 1/(2Q^2)$. We show that, even when $|Q|$ is of the order of 1, it is possible for these models to be consistent with cosmological and local gravity constraints as long as the field potential is designed in a suitable way.

We investigate the evolution of matter density perturbations for $f(R)$ and scalar-tensor models and show that model parameters as well as the strength of the coupling Q can be constrained from matter/CMB power spectra due to the enhanced growth rate of perturbations compared to the ΛCDM model.

Finally, we discuss the DGP braneworld model as a candidate for dark energy. While the late-time cosmic acceleration is possible, this model is under strong pressure from joint constraints using the data of SNLS, BAO, and the CMB shift parameter. Moreover, a ghost mode is present for such a self-accelerating universe. Thus the original DGP model is effectively ruled out from observational constraints as well as from the ghost problem.

S. Tsujikawa (✉)
Department of Physics, Faculty of Science, Tokyo University of Science, 1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan
e-mail: shinji@rs.kagu.tus.ac.jp

Tsujikawa, S.: *Modified Gravity Models of Dark Energy*. Lect. Notes Phys. **800**, 99–145 (2010)
DOI 10.1007/978-3-642-10598-2_3

1 Introduction

The origin of dark energy (DE) has been one of the most serious problems in modern cosmology [1–6]. The first step toward understanding the nature of DE is to clarify whether it is a simple cosmological constant or it originates from other sources that dynamically change in time. The dynamical DE models can be distinguished from the cosmological constant by considering the evolution of the equation of state of DE ($= w_{\rm DE}$). The scalar-field models of DE such as quintessence [7–11] and k-essence [12, 13] predict a wide variety of variations of $w_{\rm DE}$, but still the current observational data are not sufficient to provide some preference of such models over the ΛCDM model. Moreover, it is generally difficult to construct viable scalar-field models in the framework of particle physics because of a very low energy scale of the field potential.

There exists another class of dynamical DE models that modify Einstein gravity. The models that belong to this class are $f(R)$ gravity [14–17] (f is function of the Ricci scalar R), scalar-tensor theories [18–22], braneworld models [23], and Gauss–Bonnet gravity [24]. The attractive feature of these models is that cosmic acceleration can be realized without recourse to a dark energy component. If we modify gravity from General Relativity, however, there are tight constraints coming from local gravity tests as well as a number of observational constraints. Hence the restriction on modified gravity models is, in general, very tight compared to modified matter models (such as quintessence and k-essence).

For example an $f(R)$ model of the form $f(R) = R - \mu^{2(n+1)}/R^n$ was proposed to explain the late-time cosmic acceleration, but it became clear that this model is unable to satisfy local gravity constraints [25, 26] and that it also suffers from a number of problems such as the instability of density perturbations [27–30] as well as the absence of a matter-dominated epoch [31, 32]. Over the past few years there has been a burst of activity in the search for viable $f(R)$ models [27–30, 33–38]. The conditions for the viability of $f(R)$ DE models have been clarified by such extensive works, which stimulated to propose a number of workable models [39–42].

The simplest version of scalar-tensor theory is the so-called Brans–Dicke theory in which a scalar field φ is coupled to the Ricci scalar R with the Lagrangian density $\mathcal{L} = \varphi R/2 - (\omega_{\rm BD}/2\varphi)(\nabla\varphi)^2$, where $\omega_{\rm BD}$ is a so-called Brans–Dicke parameter [43]. If we allow the field potential $U(\varphi)$ in Brans–Dicke theory, the $f(R)$ models (in the metric formalism) are equivalent to this generalized Brans–Dicke theory with the Brans–Dicke parameter $\omega_{\rm BD} = 0$. If we transform the Brans–Dicke action ("Jordan frame") to the so-called Einstein frame action by a conformal transformation, the theory in the Einstein frame is equivalent to a coupled quintessence scenario [44] with a constant coupling Q satisfying the relation $1/(2Q^2) = 3 + 2\omega_{\rm BD}$ [45]. For example, the $f(R)$ gravity corresponds to the constant coupling $Q = -1/\sqrt{6}$. For the couplings $|Q|$ of the order of unity it is generally difficult to satisfy local gravity constraints unless there is some mechanism to suppress the propagation of the fifth force between the field and non-relativistic matter. As we will see in this review, it is possible for such large coupling models to be consistent with local gravity tests [39, 46, 47] through the so-called chameleon mechanism [48, 49], provided that the field has a large mass in the region of high density.

A braneworld model of dark energy was proposed by Dvali, Gabadadze, and Porrati (DGP) by embedding a 3-brane in the five-dimensional (5D) Minkowski bulk space-time [23]. In this scenario the gravitational leakage to the extra dimension leads to the modification of gravity for large distances, which causes the late-time cosmic acceleration on the 3-brane. While this self-accelerating universe is attractive, it became clear that the DGP model is disfavored by the joint observational constraints from Supernova Ia (SN Ia), baryon acoustic oscillations (BAO), and cosmic microwave background (CMB) [50–54]. Moreover, the analysis of cosmological perturbations shows that the DGP model contains a ghost mode [55–57].

In the following sections, we shall review a number of cosmological and gravitational aspects of $f(R)$ gravity, scalar-tensor theory, and the DGP braneworld model. We also discuss observational signatures of such models to distinguish them from the ΛCDM model.

2 $f(R)$ Gravity

Let us first start with the action in $f(R)$ gravity:

$$S = \frac{1}{2\kappa^2}\int \mathrm{d}^4x\sqrt{-g}f(R) + S_m(g_{\mu\nu},\Psi_m), \tag{1}$$

where $\kappa^2 = 8\pi G$ (G is a bare gravitational constant) and S_m is a matter action with matter fields Ψ_m. The field equation can be derived by varying the action (1) with respect to $g_{\mu\nu}$:

$$F(R)R_{\mu\nu}(g) - \frac{1}{2}f(R)g_{\mu\nu} - \nabla_\mu\nabla_\nu F(R) + g_{\mu\nu}\Box F(R) = \kappa^2 T_{\mu\nu}, \tag{2}$$

where $F(R) \equiv \partial f/\partial R$ and $T_{\mu\nu}$ is an energy momentum tensor of matter. The trace of (2) is given by

$$3\,\Box F(R) + F(R)R - 2f(R) = \kappa^2 T, \tag{3}$$

where $T = g^{\mu\nu}T_{\mu\nu} = -\rho + 3P$. Here ρ and P are the energy density and the pressure of the matter, respectively.

Note that there is another approach called the Palatini formalism in which $g_{\mu\nu}$ and the affine connection $\Gamma^\alpha_{\beta\gamma}$ are treated as independent variables when we vary the action (1) [58–61]. The $f(R)$ theory in the Palatini formalism gives rise to a large coupling between a scalar-field degree of freedom and ordinary matter [60–62], which is difficult to be compatible with standard particle physics. In the following we focus on the metric variational approach (so-called the metric formalism) given above. The Einstein gravity without the cosmological constant corresponds to $f(R) = R$ and $F(R) = 1$, so that the term $\Box F(R)$ in (3) vanishes. In this case we have $R = -\kappa^2 T = \kappa^2(\rho - 3P)$ and hence the Ricci scalar R is directly determined by the matter (the trace T). In modified gravity models the term $\Box F(R)$ does not

vanish in (3), which means that there is a propagating scalar degree of freedom, $\psi \equiv F(R)$. The trace equation (3) allows the dynamics of the scalar field ψ (dubbed "scalaron" [63]).

The de Sitter point corresponds to a vacuum solution at which the Ricci scalar is constant. Since $\Box F(R) = 0$ at this point, we obtain

$$F(R)R - 2f(R) = 0. \tag{4}$$

The model $f(R) = \alpha R^2$ satisfies this condition, so that it gives rise to an exact de Sitter solution [63]. In the model $f(R) = R + \alpha R^2$, the accelerated cosmic expansion ends when the term αR^2 becomes smaller than the linear term R.

2.1 Cosmological Dynamics in $f(R)$ Gravity

We first study cosmological dynamics of $f(R)$ gravity models. It is possible to carry out a general analysis without specifying the form of $f(R)$ and then to derive the conditions for the cosmological viability of $f(R)$ models. We take a Friedmann–Lemaître–Robertson–Walker (FLRW) space-time

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a(t)^2 \mathrm{d}\mathbf{x}^2, \tag{5}$$

where $a(t)$ is a scale factor. As a matter action S_m in (1) we take into account non-relativistic matter and radiation, whose energy densities satisfy the usual conservation equations $\dot{\rho}_m + 3H\rho_m = 0$ and $\dot{\rho}_r + 4H\rho_r = 0$, respectively (a dot represents a derivative with respect to cosmic time t). From (2) and (3), we obtain the following equations

$$3FH^2 = \kappa^2(\rho_m + \rho_r) + (FR - f)/2 - 3H\dot{F}, \tag{6}$$

$$2F\dot{H} = -\kappa^2\left[\rho_m + (4/3)\rho_r\right] - \ddot{F} + H\dot{F}, \tag{7}$$

where $H \equiv \dot{a}/a$ is the Hubble parameter and

$$R = 6(2H^2 + \dot{H}). \tag{8}$$

Let us introduce the following variables:

$$x_1 \equiv -\frac{\dot{F}}{HF}, \quad x_2 \equiv -\frac{f}{6FH^2}, \quad x_3 \equiv \frac{R}{6H^2}, \quad x_4 \equiv \frac{\kappa^2\rho_r}{3FH^2}, \tag{9}$$

together with the density parameters

$$\Omega_m \equiv \frac{\kappa^2\rho_m}{3FH^2} = 1 - x_1 - x_2 - x_3 - x_4, \quad \Omega_r \equiv x_4, \quad \Omega_{\rm DE} \equiv x_1 + x_2 + x_3. \tag{10}$$

It is straightforward to derive the following equations [36]:

$$\frac{dx_1}{dN} = -1 - x_3 - 3x_2 + x_1^2 - x_1 x_3 + x_4 \,, \tag{11}$$

$$\frac{dx_2}{dN} = \frac{x_1 x_3}{m} - x_2(2x_3 - 4 - x_1) \,, \tag{12}$$

$$\frac{dx_3}{dN} = -\frac{x_1 x_3}{m} - 2x_3(x_3 - 2) \,, \tag{13}$$

$$\frac{dx_4}{dN} = -2x_3 x_4 + x_1 \, x_4, \tag{14}$$

where $N = \log(a)$ and

$$m \equiv \frac{d\log F}{d\log R} = \frac{R f_{,RR}}{f_{,R}}, \tag{15}$$

$$r \equiv -\frac{d\log f}{d\log R} = -\frac{R f_{,R}}{f} = \frac{x_3}{x_2}. \tag{16}$$

From (16), one can express R as a function of x_3/x_2. Since m is a function of R, it follows that m is a function of r, i.e., $m = m(r)$. The ΛCDM model, $f(R) = R - 2\Lambda$, corresponds to $m = 0$. Hence the quantity m characterizes the deviation from the ΛCDM model.

The effective equation of state of the system is given by

$$w_{\rm eff} = -(2x_3 - 1)/3. \tag{17}$$

In the absence of radiation ($x_4 = 0$), the fixed points for the dynamical system (11), (12), (13), and (14) are

$$P_1{:}(x_1,x_2,x_3) = (0,-1,2), \quad \Omega_m = 0, \quad w_{\rm eff} = -1, \tag{18}$$

$$P_2{:}(x_1,x_2,x_3) = (-1,0,0), \quad \Omega_m = 2, \quad w_{\rm eff} = 1/3, \tag{19}$$

$$P_3{:}(x_1,x_2,x_3) = (1,0,0), \quad \Omega_m = 0, \quad w_{\rm eff} = 1/3, \tag{20}$$

$$P_4{:}(x_1,x_2,x_3) = (-4,5,0), \quad \Omega_m = 0, \quad w_{\rm eff} = 1/3, \tag{21}$$

$$P_5{:}(x_1,x_2,x_3) = \left(\frac{3m}{1+m}, -\frac{1+4m}{2(1+m)^2}, \frac{1+4m}{2(1+m)}\right), \tag{22}$$

$$\Omega_m = 1 - \frac{m(7+10m)}{2(1+m)^2}, \quad w_{\rm eff} = -\frac{m}{1+m}, \tag{23}$$

$$P_6{:}(x_1,x_2,x_3) = \left(\frac{2(1-m)}{1+2m}, \frac{1-4m}{m(1+2m)}, -\frac{(1-4m)(1+m)}{m(1+2m)}\right),$$

$$\Omega_m = 0, \quad w_{\rm eff} = \frac{2-5m-6m^2}{3m(1+2m)}. \tag{24}$$

The points P_5 and P_6 are on the line $m(r) = -r - 1$ in the (r,m) plane.

Only the point P_5 can be responsible for the matter-dominated epoch ($\Omega_m \simeq 1$ and $w_{\rm eff} \simeq 0$), which can be realized for m close to 0. In the (r,m) plane this point

exists around $(r,m) = (-1,0)$. Either the point P_1 or P_6 can lead to the late-time cosmic acceleration. The former corresponds to a de Sitter point ($w_{\rm eff} = -1$) with $r = -2$, in which case the condition (4) is in fact satisfied. The point P_6 is able to give the accelerated expansion depending on the values of m. In the following we shall focus on the case in which the matter point P_5 is followed by the de Sitter point P_1.

The stability of the fixed points is known by considering small perturbations δx_i ($i = 1,2,3$) around them [36]. For the point P_5 the eigenvalues for the 3×3 Jacobian matrix of perturbations are

$$3(1+m_5'),\ \frac{-3m_5 \pm \sqrt{m_5(256m_5^3+160m_5^2-31m_5-16)}}{4m_5(m_5+1)}, \tag{25}$$

where $m_5 \equiv m(r_5)$ and $m_5' \equiv \frac{{\rm d}m}{{\rm d}r}(r_5)$, with $r_5 \approx -1$. In the limit $|m_5| \ll 1$, the latter two eigenvalues reduce to $-3/4 \pm \sqrt{-1/m_5}$. The $f(R)$ models with $m_5 < 0$ show a divergence of the eigenvalues as $m_5 \to -0$, in which case the system cannot remain for a long time around the point P_5. For example, the model $f(R) = R - \alpha/R^n$ with $n > 0$ and $\alpha > 0$ falls into this category. On the other hand, if $0 < m_5 < 0.327$, the latter two eigenvalues in (25) are complex with negative real parts. Then, provided that $m_5' > -1$, the point P_5 corresponds to a saddle point with a damped oscillation. Hence the universe can evolve toward the point P_5 from the radiation era and leave for the late-time acceleration. Then the condition for the existence of the saddle matter era is

$$m(r) \approx +0, \quad \frac{{\rm d}m}{{\rm d}r} > -1, \quad \text{at} \quad r = -1. \tag{26}$$

The first condition implies that the $f(R)$ models need to be close to the ΛCDM model during the matter era.

The eigenvalues for the Jacobian matrix of perturbations about the point P_1 are

$$-3, \quad -\frac{3}{2} \pm \frac{\sqrt{25-16/m_1}}{2}, \tag{27}$$

where $m_1 = m(r = -2)$. This shows that the condition for the stability of the de Sitter point P_1 is

$$0 < m(r = -2) \leq 1. \tag{28}$$

The trajectories that start from the saddle matter point P_5 satisfying the condition (26) and then approach the stable de Sitter point P_1 satisfying the condition (28) are cosmologically viable.

Let us consider a couple of viable $f(R)$ models in the (r,m) plane. The ΛCDM model, $f(R) = R - 2\Lambda$, corresponds to $m = 0$, in which case the trajectory is a straight line from P_5: $(r,m) = (-1,0)$ to P_1: $(r,m) = (-2,0)$. The trajectory (ii) in Fig. 1 represents the model $f(R) = (R^b - \Lambda)^c$ [64], which corresponds to the straight

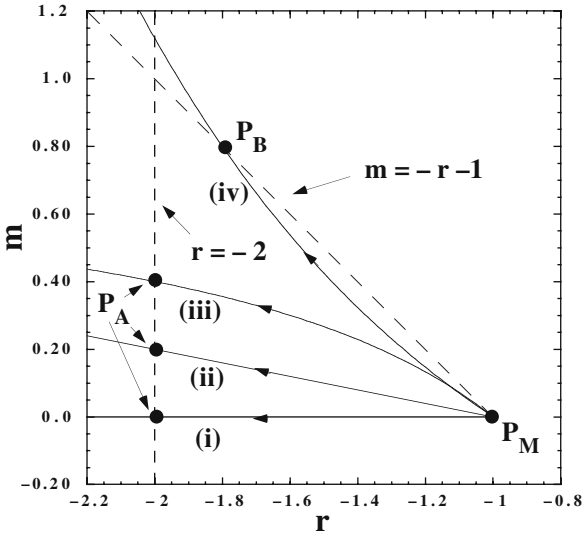

Fig. 1 Four trajectories in the (r,m) plane. Each trajectory corresponds to the models: (i) ΛCDM, (ii) $f(R) = (R^b - \Lambda)^c$, (iii) $f(R) = R - \alpha R^n$ with $\alpha > 0, 0 < n < 1$, and (iv) $m(r) = -C(r+1)(r^2 + ar + b)$. Here P_M, P_A, and P_B are the matter point P_5, the de Sitter point P_1, and the accelerated point P_6, respectively

line $m(r) = [(1-c)/c]r + b - 1$ in the (r,m) plane. The existence of a saddle matter epoch requires the condition $c \geq 1$ and $bc \approx 1$. The trajectory (iii) represents the model [36, 37]

$$f(R) = R - \alpha R^n \qquad (\alpha > 0, 0 < n < 1), \tag{29}$$

which corresponds to the curve $m = n(1+r)/r$. The trajectory (iv) in Fig. 1 shows the model $m(r) = -C(r+1)(r^2 + ar + b)$, in which case the late-time accelerated attractor is the point P_6 with $(\sqrt{3} - 1)/2 < m < 1$.

In [36] it was shown that the variable m needs to be close to 0 during the radiation-dominated epoch as well. Hence the viable $f(R)$ models are close to the ΛCDM model, $f(R) = R - 2\Lambda$, in the region $R \gg R_0$ (where R_0 is the present cosmological Ricci scalar). The Ricci scalar given in (8) remains positive from the radiation era to the present epoch, as long as it does not oscillate. Note that we require the condition $f_{,R} > 0$ to avoid anti-gravity. Then the condition $m > 0$ for the presence of the matter era translates into $f_{,RR} > 0$. The model $f(R) = R - \alpha/R^n$ ($\alpha > 0, n > 0$) is not viable because the condition $f_{,RR} > 0$ is violated.

In order to derive the equation of state of dark energy to confront with SN Ia observations for the cosmologically viable models, we rewrite (6) and (7) as follows:

$$3AH^2 = \kappa^2 (\rho_m + \rho_r + \rho_{\rm DE}), \tag{30}$$

$$-2A\dot{H} = \kappa^2 \left[\rho_m + (4/3)\rho_r + \rho_{\rm DE} + P_{\rm DE}\right], \tag{31}$$

where A is some constant and

$$\kappa^2\rho_{\rm DE} \equiv (1/2)(FR-f) - 3H\dot{F} + 3H^2(A-F), \tag{32}$$

$$\kappa^2 P_{\rm DE} \equiv \ddot{F} + 2H\dot{F} - (1/2)(FR-f) - (3H^2+2\dot{H})(A-F). \tag{33}$$

Defining $\rho_{\rm DE}$ and $P_{\rm DE}$ in the above way, we find that these satisfy the usual continuity equation

$$\dot{\rho}_{\rm DE} + 3H(\rho_{\rm DE} + P_{\rm DE}) = 0. \tag{34}$$

The dark energy equation of state, $w_{\rm DE} \equiv P_{\rm DE}/\rho_{\rm DE}$, is directly related to the one used in SN Ia observations. From (30) and (31) it is given by

$$w_{\rm DE} = -\frac{2A\dot{H} + 3AH^2 + \kappa^2\rho_r/3}{3AH^2 - \kappa^2(\rho_m+\rho_r)} \simeq \frac{w_{\rm eff}}{1-(F/A)\Omega_m}, \tag{35}$$

where the last approximate equality in (35) is valid in the regime where the radiation density ρ_r is negligible relative to the matter density. The viable $f(R)$ models approach the ΛCDM model in the past, i.e., $F \to 1$ as $R \to \infty$. In order to reproduce the standard matter era for $z \gg 1$, we can choose $A = 1$ in (30) and (31). Another possible choice is $A = F_0$, where F_0 is the present value of F. This choice is suitable if the deviation of F_0 from 1 is small (as in the scalar-tensor theory with a massless scalar field [65]). In both cases the equation of state $w_{\rm DE}$ can be smaller than -1 before reaching the de Sitter attractor [39, 64, 66]. Thus $f(R)$ gravity models give rise to a phantom equation of state without violating stability conditions of the system.

2.2 Local Gravity Constraints on $f(R)$ Gravity Models

It is required that $f(R)$ gravity models satisfy local gravity constraints as well as the conditions for the cosmological viability. In an environment of high density such as Earth or Sun, the Ricci scalar R is much larger than the background cosmological value R_0. In such a non-linear regime the effect of the chameleon mechanism is crucially important. It is possible for $f(R)$ models to be consistent with local gravity constraints under the chameleon mechanism.

In order to discuss the chameleon mechanism in $f(R)$ gravity, it is convenient to transform the action (1) to the so-called Einstein frame action via the conformal transformation [67]:

$$\tilde{g}_{\mu\nu} = \Omega^2 g_{\mu\nu}, \tag{36}$$

where Ω is a so-called conformal factor. In the Einstein frame the Lagrangian includes a linear term in $\tilde{R}$ (the tilde represents quantities in the Einstein frame). The Ricci scalars in the two frames have the following relation:

$$R = \Omega^2(\tilde{R} + 6\tilde{\Box}\omega - 6\tilde{g}^{\mu\nu}\omega_{,\mu}\omega_{,\nu}), \tag{37}$$

where

$$\omega_{,\mu} \equiv \frac{\partial_\mu \Omega}{\Omega}, \qquad \tilde{\Box}\omega \equiv \frac{1}{\sqrt{-\tilde{g}}}\partial_\mu(\sqrt{-\tilde{g}}\,\tilde{g}^{\mu\nu}\partial_\nu\omega). \tag{38}$$

The action (1) can be written as

$$S = \int \mathrm{d}^4x\sqrt{-g}\left(\frac{1}{2\kappa^2}FR - U\right) + S_m(g_{\mu\nu}, \Psi_m), \tag{39}$$

where

$$U = (RF - f)/(2\kappa^2). \tag{40}$$

Using (37) and the relation $\sqrt{-g} = \Omega^{-4}\sqrt{-\tilde{g}}$, the action (39) is transformed to be

$$S = \int \mathrm{d}^4x\sqrt{-\tilde{g}}\left[\frac{1}{2\kappa^2}F\Omega^{-2}(\tilde{R} + 6\tilde{\Box}\omega - 6\tilde{g}^{\mu\nu}\omega_{,\mu}\omega_{,\nu}) - \Omega^{-4}U\right] + S_m(g_{\mu\nu}, \Psi_m). \tag{41}$$

We obtain a linear action in $\tilde{R}$ for the choice

$$\Omega^2 = F. \tag{42}$$

We also introduce a new scalar field ϕ defined by

$$\phi \equiv \sqrt{3/2}\ln F. \tag{43}$$

Since $\Omega = \sqrt{F}$ and $\omega_{,\mu} = \Omega_{,\mu}/\Omega$, it follows that $\omega_{,\mu} = (1/\sqrt{6})\kappa\phi_{,\mu}$. The integral, $\int \mathrm{d}^4x\sqrt{-\tilde{g}}\,\tilde{\Box}\omega$, vanishes due to the Gauss's theorem by using (38). Then the action in the Einstein frame is

$$S_E = \int \mathrm{d}^4x\sqrt{-\tilde{g}}\left[\frac{1}{2\kappa^2}\tilde{R} - \frac{1}{2}\tilde{g}^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - V(\phi)\right] + S_m(g_{\mu\nu}, \Psi_m), \tag{44}$$

where

$$V(\phi) = (RF - f)/(2\kappa^2F^2). \tag{45}$$

In the following we use the unit $\kappa^2 = 1$. In the Einstein frame the scalar field ϕ directly couples with non-relativistic matter. The strength of this coupling depends on the field dependence of the conformal factor Ω in (36). Let us define the coupling Q as

$$Q \equiv -\frac{\Omega_{,\phi}}{\Omega} = -\frac{F_{,\phi}}{2F} = -\frac{1}{\sqrt{6}}, \tag{46}$$

whose strength is of the order of unity in $f(R)$ gravity. In the absence of the field potential $V(\phi)$ it is not possible to satisfy local gravity constraints because the field propagates freely with a large coupling Q. Since a potential $V(\phi)$ with a gravitational origin is present in $f(R)$ gravity, local gravity tests can be escaped through the chameleon mechanism [48, 49], provided that the form of $f(R)$ is appropriately chosen [35, 39, 46, 47].

In the following we shall discuss how the chameleon mechanism can make the $f(R)$ models consistent with local gravity constraints. In a spherically symmetric space-time under the weak gravitational background (i.e., neglecting the back reaction of gravitational potentials), the variation of the action (44) with respect to the scalar field ϕ gives

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}\tilde{r}^2}+\frac{2}{\tilde{r}}\frac{\mathrm{d}\phi}{\mathrm{d}\tilde{r}}=\frac{\mathrm{d}V_{\rm eff}}{\mathrm{d}\phi}, \tag{47}$$

where $\tilde{r}$ is a distance from the center of symmetry and

$$V_{\rm eff}(\phi)=V(\phi)+e^{Q\phi}\rho^*. \tag{48}$$

Here ρ^* is a conserved quantity in the Einstein frame, which is related to the energy density $\tilde{\rho}$ in the Jordan frame via the relation $\rho^*=e^{3Q\phi}\tilde{\rho}$.

We assume that a spherically symmetric body has a constant density $\rho^*=\rho_A$ inside the body ($\tilde{r}<\tilde{r}_c$) and that the energy density outside the body ($\tilde{r}>\tilde{r}_c$) is $\rho^*=\rho_B$. The mass M_c of the body and the gravitational potential Φ_c at the radius $\tilde{r}_c$ are given by $M_c=(4\pi/3)\tilde{r}_c^3\rho_A$ and $\Phi_c=M_c/8\pi\tilde{r}_c$, respectively. The effective potential $V_{\rm eff}(\phi)$ has two minima at the field values ϕ_A and ϕ_B satisfying $V'_{\rm eff}(\phi_A)=0$ and $V'_{\rm eff}(\phi_B)=0$, respectively. The former corresponds to the region with a high density that gives rise to a heavy mass squared $m_A^2\equiv V''_{\rm eff}(\phi_A)$, whereas the latter to the lower density region with a lighter mass squared $m_B^2\equiv V''_{\rm eff}(\phi_B)$. Note that when we consider the "dynamics" of the field ϕ according to (47), we need to consider the effective potential $(-V_{\rm eff})$ so that it has two *maxima* at $\phi=\phi_A$ and $\phi=\phi_B$.

We impose the following boundary conditions:

$$\frac{\mathrm{d}\phi}{\mathrm{d}\tilde{r}}(\tilde{r}=0)=0,\quad \phi(\tilde{r}\to\infty)=\phi_B. \tag{49}$$

The field ϕ is at rest at $\tilde{r}=0$ and begins to roll down the potential when the matter-coupling term $Q\rho_A e^{Q\phi}$ becomes important at a radius $\tilde{r}_1$ in (47). As long as $\tilde{r}_1$ is close to $\tilde{r}_c$ so that $\Delta\tilde{r}_c\equiv\tilde{r}_c-\tilde{r}_1\ll\tilde{r}_c$, the body has a thin shell inside the body. The field acquires a sufficient kinetic energy in the thin-shell regime ($\tilde{r}_1<\tilde{r}<\tilde{r}_c$) and hence the field climbs up the potential hill outside the body ($\tilde{r}>\tilde{r}_c$).

The field profile can be obtained by matching the solutions of (47) at the radius $\tilde{r}=\tilde{r}_1$ and $\tilde{r}=\tilde{r}_c$. Neglecting the mass term m_B, the thin-shell field profile outside the body is given by [68]

$$\phi(r) = \phi_B - 2Q_{\rm eff}\frac{GM_c}{\tilde{r}}, \tag{50}$$

where

$$Q_{\rm eff} \simeq 3Q\varepsilon_{\rm th}, \qquad \varepsilon_{\rm th} \equiv \frac{\phi_B - \phi_A}{6Q\Phi_c}. \tag{51}$$

Here $\varepsilon_{\rm th}$ is called the thin-shell parameter. Under the conditions $\Delta\tilde{r}_c/\tilde{r}_c \ll 1$ and $1/(m_A\tilde{r}_c) \ll 1$, the thin-shell parameter is approximately given by [68]

$$\varepsilon_{\rm th} \simeq \frac{\Delta\tilde{r}_c}{\tilde{r}_c} + \frac{1}{m_A\tilde{r}_c}. \tag{52}$$

As long as $\varepsilon_{\rm th} \ll 1$, the amplitude of the effective coupling $Q_{\rm eff}$ can be much smaller than 1. Hence it is possible for the large coupling models ($|Q| = \mathcal{O}(1)$) to be consistent with local gravity experiments if the body has a thin shell. In the original papers of Khoury and Weltman [48, 49] the thin-shell solution was derived by assuming that the field is frozen in the region $0 < \tilde{r} < \tilde{r}_1$. In this case the thin-shell parameter is given by $\varepsilon_{\rm th} \simeq \Delta\tilde{r}_c/\tilde{r}_c$, which is different from (52). However, this difference is not important because the condition $\Delta\tilde{r}_c/\tilde{r}_c \gg 1/(m_A\tilde{r}_c)$ is satisfied for most of viable models [68].

To be concrete let us consider the constraint on the thin-shell parameter from the possible violation of the equivalence principle (EP). The tightest bound comes from the solar system tests of weak EP using the free-fall acceleration of Moon ($a_{\rm Moon}$) and Earth ($a_\oplus$) toward Sun [49]. The experimental bound on the difference of two accelerations is given by [69]

$$2\frac{|a_{\rm Moon} - a_\oplus|}{a_{\rm Moon} + a_\oplus} < 10^{-13}. \tag{53}$$

Under the conditions that Earth, Sun, and Moon have thin shells, the field profiles outside the bodies are given by (50) with the replacement of corresponding quantities. The acceleration induced by a fifth force with the field profile $\phi(r)$ and the effective coupling $Q_{\rm eff}$ is $a^{\rm fifth} = |Q_{\rm eff}\nabla\phi(r)|$. Using the thin-shell parameter $\varepsilon_{{\rm th},\oplus}$ for Earth, the accelerations $a_\oplus$ and $a_{\rm Moon}$ toward Sun (mass $M_\odot$) are

$$a_\oplus \simeq \frac{GM_\odot}{r^2}\left[1 + 18Q^2\varepsilon_{{\rm th},\oplus}^2\frac{\Phi_\oplus}{\Phi_\odot}\right], \tag{54}$$

$$a_{\rm Moon} \simeq \frac{GM_\odot}{r^2}\left[1 + 18Q^2\varepsilon_{{\rm th},\oplus}^2\frac{\Phi_\oplus^2}{\Phi_\odot\Phi_{\rm Moon}}\right], \tag{55}$$

where $\Phi_\odot \simeq 2.1 \times 10^{-6}$, $\Phi_\oplus \simeq 7.0 \times 10^{-10}$, and $\Phi_{\rm Moon} \simeq 3.1 \times 10^{-11}$ are the gravitational potentials of Sun, Earth, and Moon, respectively. Hence the condition (53) translates into

$$\varepsilon_{\rm th,\oplus} < \frac{8.8\times 10^{-7}}{|Q|}. \tag{56}$$

Since the condition $|\phi_B| \gg |\phi_A|$ is satisfied for viable $f(R)$ models we study in the following, we have $\varepsilon_{\rm th,\oplus} \simeq \phi_B/6Q\Phi_\oplus$ from (51). Then the condition (56) translates into

$$|\phi_{B,\oplus}| < 3.7\times 10^{-15}. \tag{57}$$

In the following, we list some viable $f(R)$ models that can be consistent with local gravity tests as well as cosmological constraints:

$$\text{(A)}\, f(R) = R - \mu R_c \frac{(R/R_c)^{2n}}{(R/R_c)^{2n}+1} \qquad \text{with } n>0 \text{ and } R_c>0, \tag{58}$$

$$\text{(B)}\, f(R) = R - \mu R_c \left[1-\left(1+R^2/R_c^2\right)^{-n}\right] \quad \text{with } n>0 \text{ and } R_c>0, \tag{59}$$

$$\text{(C)}\, f(R) = R - \mu R_c \tanh(R/R_c) \qquad \text{with } R_c>0. \tag{60}$$

The models (A), (B), and (C) have been proposed in [39, 40], and [42], respectively. A model similar to (C) has been also proposed in [41]. In the models (A) and (B) the function $f(R)$ asymptotically behaves as

$$f(R) \simeq R - \mu R_c[1-(R^2/R_c^2)^{-n}] \quad \text{for} \quad R \gg R_c. \tag{61}$$

In the model (C) the function $f(R)$ rapidly approaches $f(R) \to R-\mu R_c$ in the region $R \gg R_c$.

Let us consider local gravity constraints on the $f(R)$ models given in (58) and (59). In the region of high density where local gravity experiments are carried out, it is sufficient to use the asymptotic form given in (61). In order for these models to be responsible for the present cosmic acceleration, R_c is roughly the same order as the cosmological Ricci scalar R_0 today for μ and n of the order of unity. For the functional form (61), we have the following relations:

$$F = e^{2\phi/\sqrt{6}} = 1 - 2n\mu(R/R_c)^{-(2n+1)}, \tag{62}$$

$$V_{\rm eff}(\phi) \simeq \frac{1}{2}\mu R_c e^{-4\phi/\sqrt{6}}\left[1-(2n+1)\left(\frac{-\phi}{\sqrt{6}n\mu}\right)^{2n/(2n+1)}\right] + \rho^* e^{-\phi/\sqrt{6}}. \tag{63}$$

Inside and outside the body the effective potential (63) has minima at

$$\phi_A \simeq -\sqrt{6}n\mu(R_c/\rho_A)^{2n+1}, \quad \phi_B \simeq -\sqrt{6}n\mu(R_c/\rho_B)^{2n+1}, \tag{64}$$

which satisfies $|\phi_B| \gg |\phi_A|$ for $\rho_A \gg \rho_B$.

The bound (57) translates into

$$\frac{n\mu}{x_1^{2n+1}}\left(\frac{R_1}{\rho_B}\right)^{2n+1} < 1.5 \times 10^{-15}, \tag{65}$$

where x_1 is defined by $x_1 \equiv R_1/R_c$. Let us consider the case in which the Lagrangian density is given by (61) for $R \geq R_1$. If we use the original models of Hu and Sawicki [39] and Starobinsky [40], then there are small modifications for the estimation of R_1, but this change is insignificant when we place constraints on model parameters.

The de Sitter point for the model (61) corresponds to $\mu = x_1^{2n+1}/[2(x_1^{2n} - n - 1)]$. Substituting this relation into (65), it follows that

$$\frac{n}{2(x_1^{2n} - n - 1)}\left(\frac{R_1}{\rho_B}\right)^{2n+1} < 1.5 \times 10^{-15}. \tag{66}$$

For the stability of the de Sitter point we require that $m(R_1) < 1$, which translates into the condition $x_1^{2n} > 2n^2 + 3n + 1$. Hence the term $n/[2(x_1^{2n} - n - 1)]$ in (66) is smaller than 0.25 for $n > 0$.

Now it is possible to make an approximation that R_1 and ρ_B are of the orders of the present cosmological density 10^{-29} g/cm^3 and the baryonic/dark matter density 10^{-24} g/cm^3 in our galaxy, respectively. From (66) we obtain the constraint [46]

$$n > 0.9. \tag{67}$$

Thus n does not need to be much larger than unity. Under the condition (67) one can see an appreciable deviation from the ΛCDM model cosmologically as R decreases to the order of R_c.

Thus, we have shown that the models (58) and (59) are consistent with local gravity tests for $n > 0.9$. The deviation from the ΛCDM model appears when R decreases to the order of R_c. The model (60) also shows similar behavior. If we consider the model (29), it was shown in [46] that the bound (57) gives the constraint $n < 3 \times 10^{-10}$. This means that the deviation from the ΛCDM model is very small. The models (58) and (59) are carefully constructed to satisfy local gravity constraints, while at the same time the deviation from the ΛCDM model appears even for $n = \mathcal{O}(1)$.

3 Scalar-Tensor Theories

There is another class of modified gravity called scalar-tensor theories in which the Ricci scalar R is coupled to a scalar field φ. The simplest example is the so-called Brans–Dicke theory with the action

$$S = \int \mathrm{d}^4x\sqrt{-g}\left[\frac{1}{2}\varphi R - \frac{\omega_{\rm BD}}{2\varphi}(\nabla\varphi)^2 - U(\varphi)\right] + S_m(g_{\mu\nu}, \Psi_m), \tag{68}$$

where $\omega_{\rm BD}$ is the Brans–Dicke parameter, $U(\varphi)$ is the field potential, and S_m is a matter Lagrangian that depends on the metric $g_{\mu\nu}$ and matter fields Ψ_m. The original Brans–Dicke theory [43] does not have the field potential. As we will see below, the $f(R)$ gravity we have discussed in the previous section is equivalent to the Brans–Dicke theory with $\omega_{\rm BD}=0$.

The general action for scalar-tensor theories can be written as

$$S=\int {\rm d}^4x\sqrt{-g}\left[\frac{1}{2}f(\varphi,R)-\frac{1}{2}\zeta(\varphi)(\nabla\varphi)^2\right]+S_m(g_{\mu\nu},\Psi_m), \tag{69}$$

where f is a general function of the scalar field φ and the Ricci scalar R and ζ is a function of φ. We choose the unit $\kappa^2=1$. The action (69) includes a wide variety of theories such as $f(R)$ gravity ($f(\varphi,R)=f(R)$, $\zeta=0$) and Brans–Dicke theory ($f=\varphi R$ and $\zeta=\omega_{\rm BD}/\varphi$). The action (69) can be transformed to that in the Einstein frame under the conformal transformation (36) with the choice

$$\Omega^2=F\equiv\frac{\partial f}{\partial R}, \tag{70}$$

where F is positive as long as gravity is attractive.

We consider theories of the type

$$f(\varphi,R)=F(\varphi)R-2U(\varphi), \tag{71}$$

in which case the conformal factor Ω depends on φ only. Under the conformal transformation (70) the action in the Einstein frame is given by

$$S_E=\int {\rm d}^4x\sqrt{-\tilde{g}}\left[\frac{1}{2}\tilde{R}-\frac{1}{2}(\tilde{\nabla}\phi)^2-V(\phi)\right]+S_m(\tilde{g}_{\mu\nu}F^{-1},\Psi_m), \tag{72}$$

where

$$V=U/F^2. \tag{73}$$

Note that we have introduced a new scalar field ϕ in order to make the kinetic term canonical:

$$\phi\equiv\int {\rm d}\varphi\sqrt{\frac{3}{2}\left(\frac{F_{,\varphi}}{F}\right)^2+\frac{\zeta}{F}}. \tag{74}$$

We define the coupling between dark energy and non-relativistic matter:

$$Q\equiv-\frac{F_{,\phi}}{2F}=-\frac{F_{,\varphi}}{F}\left[\frac{3}{2}\left(\frac{F_{,\varphi}}{F}\right)^2+\frac{\zeta}{F}\right]^{-1/2}. \tag{75}$$

Recall that in $f(R)$ gravity we have that $Q = -1/\sqrt{6}$. If Q is a constant, the following relations hold from (74) and (75):

$$F = e^{-2Q\phi}, \quad \zeta = (1 - 6Q^2)F\left(\frac{\mathrm{d}\phi}{\mathrm{d}\varphi}\right)^2. \tag{76}$$

Then the action (69) in the Jordan frame reduces to [45]

$$S = \int \mathrm{d}^4x\sqrt{-g}\left[\frac{1}{2}F(\phi)R - \frac{1}{2}(1 - 6Q^2)F(\phi)(\nabla\phi)^2 - U(\phi)\right] + S_m(g_{\mu\nu}, \Psi_m). \tag{77}$$

In the limit $Q \to 0$, the action (77) reduces to the one for a minimally coupled scalar field ϕ with the potential $U(\phi)$. The transformation of the Jordan frame action (77) via a conformal transformation $\tilde{g}_{\mu\nu} = F(\phi)g_{\mu\nu}$ gives rise to the Einstein frame action (72) with a constant coupling Q. Note that this is equivalent to the action (44) with $\tilde{g}_{\mu\nu} = e^{-2Q\phi}g_{\mu\nu}$.

One can compare (77) with the action (68) in Brans–Dicke theory. Setting $\varphi = F = e^{-2Q\phi}$, one finds that two actions are equivalent if the parameter $\omega_{\rm BD}$ is related to Q via the relation [45]

$$3 + 2\omega_{\rm BD} = \frac{1}{2Q^2}. \tag{78}$$

Using this relation, we find that the General Relativistic limit ($\omega_{\rm BD} \to \infty$) corresponds to the vanishing coupling ($Q \to 0$). Since $Q = -1/\sqrt{6}$ in $f(R)$ gravity, this corresponds to the Brans–Dicke parameter $\omega_{\rm BD} = 0$ [25].

In the following we shall study the cosmological dynamics and local gravity constraints on the constant coupling models based on the action (77) with $F(\phi) = e^{-2Q\phi}$.

3.1 Cosmological Dynamics

We study the cosmological dynamics for the Jordan frame action (77) in the presence of a non-relativistic fluid with energy density ρ_m and a radiation fluid with energy density ρ_r. The Jordan frame is regarded as a physical frame due to the usual conservation of non-relativistic matter ($\rho_m \propto a^{-3}$). In the flat FLRW background the variation of the action (77) with respect to $g_{\mu\nu}$ and ϕ gives the following equations of motion:

$$3FH^2 = (1/2)(1 - 6Q^2)F\dot{\phi}^2 + U - 3H\dot{F} + \rho_m + \rho_r, \tag{79}$$

$$2F\dot{H} = -(1 - 6Q^2)F\dot{\phi}^2 - \ddot{F} + H\dot{F} - \rho_m - (4/3)\rho_r, \tag{80}$$

$$(1 - 6Q^2)F\left[\ddot{\phi} + 3H\dot{\phi} + (\dot{F}/2F)\dot{\phi}\right] + U_{,\phi} + QFR = 0. \tag{81}$$

Let us introduce the following variables:

$$x_1 \equiv \frac{\dot{\phi}}{\sqrt{6}H}, \quad x_2 \equiv \frac{1}{H}\sqrt{\frac{U}{3F}}, \quad x_3 \equiv \frac{1}{H}\sqrt{\frac{\rho_r}{3F}}, \tag{82}$$

and

$$\Omega_m \equiv \frac{\rho_m}{3FH^2}, \qquad \Omega_{\rm rad} \equiv x_3^2, \qquad \Omega_{\rm DE} \equiv (1-6Q^2)x_1^2 + x_2^2 + 2\sqrt{6}Qx_1. \tag{83}$$

These satisfy the relation $\Omega_m + \Omega_{\rm rad} + \Omega_{\rm DE} = 1$ from (79). On using (79), (80), and (81), we obtain the differential equations for x_1, x_2, and x_3:

$$\frac{{\rm d}x_1}{{\rm d}N} = \frac{\sqrt{6}}{2}(\lambda x_2^2 - \sqrt{6}x_1) + \frac{\sqrt{6}Q}{2}\left[(5-6Q^2)x_1^2 + 2\sqrt{6}Qx_1 - 3x_2^2 + x_3^2 - 1\right] - x_1\frac{\dot{H}}{H^2}, \tag{84}$$

$$\frac{{\rm d}x_2}{{\rm d}N} = \frac{\sqrt{6}}{2}(2Q-\lambda)x_1x_2 - x_2\frac{\dot{H}}{H^2}, \tag{85}$$

$$\frac{{\rm d}x_3}{{\rm d}N} = \sqrt{6}Qx_1x_3 - 2x_3 - x_3\frac{\dot{H}}{H^2}, \tag{86}$$

where $\lambda \equiv -U_{,\phi}/U$ and

$$\frac{\dot{H}}{H^2} = -\frac{1-6Q^2}{2}\left(3 + 3x_1^2 - 3x_2^2 + x_3^2 - 6Q^2x_1^2 + 2\sqrt{6}Qx_1\right) + 3Q(\lambda x_2^2 - 4Q), \tag{87}$$

from which the effective equation of state of the system can be found by $w_{\rm eff} = -1 - 2\dot{H}/3H^2$.

If λ is a constant, one can derive the fixed points of the system (84), (85), and (86) in the absence of radiation ($x_3 = 0$) [45]:

- (a)

$$(x_1, x_2) = \left(\frac{\sqrt{6}Q}{3(2Q^2-1)}, 0\right), \quad \Omega_m = \frac{3-2Q^2}{3(1-2Q^2)^2}, \quad w_{\rm eff} = \frac{4Q^2}{3(1-2Q^2)}. \tag{88}$$

- (b)

$$(x_1, x_2) = \left(\frac{1}{\sqrt{6}Q \pm 1}, 0\right), \quad \Omega_m = 0, \quad w_{\rm eff} = \frac{3 \mp \sqrt{6}Q}{3(1 \pm \sqrt{6}Q)}. \tag{89}$$

- (c)

$$(x_1,x_2) = \left(\frac{\sqrt{6}(4Q-\lambda)}{6(4Q^2-Q\lambda-1)}, \left[\frac{6-\lambda^2+8Q\lambda-16Q^2}{6(4Q^2-Q\lambda-1)^2} \right]^{1/2} \right), \ \Omega_m = 0,$$
$$w_{\rm eff} = -\frac{20Q^2-9Q\lambda-3+\lambda^2}{3(4Q^2-Q\lambda-1)}. \tag{90}$$

- (d)

$$(x_1,x_2) = \left(\frac{\sqrt{6}}{2\lambda}, \sqrt{\frac{3+2Q\lambda-6Q^2}{2\lambda^2}} \right), \ \Omega_m = 1 - \frac{3-12Q^2+7Q\lambda}{\lambda^2}, \ w_{\rm eff} = -\frac{2Q}{\lambda}. \tag{91}$$

- (e)

$$(x_1,x_2) = (0,1), \quad \Omega_m = 0, \quad w_{\rm eff} = -1. \tag{92}$$

The point (e) corresponds to the de Sitter point. This exists only for $\lambda = 4Q$, which can be confirmed by setting $\dot{\phi} = 0$ in (79), (80), and (81). This is the special case of the scalar-field dominated point (c).

We first study the case of non-zero values of Q with constant λ, i.e., for the exponential potential $U(\phi) = U_0 e^{-\lambda\phi}$. We do not consider the special case of $\lambda = 4Q$. The matter-dominated era can be realized either by the point (a) or by the point (d). If the point (a) is responsible for the matter era, the condition $Q^2 \ll 1$ is required. We then have $\Omega_m \simeq 1 + 10Q^2/3 > 1$ and $w_{\rm eff} \simeq 4Q^2/3$. When $Q^2 \ll 1$ the scalar-field-dominated point (c) yields an accelerated expansion of the universe provided that $-\sqrt{2} + 4Q < \lambda < \sqrt{2} + 4Q$. Under these conditions the point (a) is followed by the late-time cosmic acceleration. The scaling solution (d) can give rise to the equation of state, $w_{\rm eff} \simeq 0$ for $|Q| \ll |\lambda|$. In this case, however, the condition $w_{\rm eff} < -1/3$ for the point (c) gives $\lambda^2 < 2$. Then the energy fraction of the pressureless matter for the point (d) does not satisfy the condition $\Omega_m \simeq 1$. From the above discussion the viable cosmological trajectory for constant λ corresponds to the sequence from the point (a) to the scalar-field-dominated point (c) under the conditions $Q^2 \ll 1$ and $-\sqrt{2} + 4Q < \lambda < \sqrt{2} + 4Q$.

We shall proceed to the case where λ varies with time. The fixed points derived above in the case of constant λ can be regarded as the "instantaneous" fixed points, provided that the timescale of the variation of λ is smaller than that of the cosmic expansion. It is then possible that the matter era is realized by the point (d) with $|Q| \ll |\lambda|$ and that the solutions finally approach either the de Sitter point (e) with $\lambda = 4Q$ or the accelerated point (c).

In the following we focus on the case in which the matter solution (d) is followed by the de Sitter solution (e). In order to study the stability of the point (e) we define a variable $x_4 \equiv F$, satisfying the following equation:

$$\frac{dx_4}{dN} = -2\sqrt{6}Qx_1x_4. \tag{93}$$

Considering the 3×3 matrix for perturbations δx_1, δx_2, and δx_4 around the point (e), we obtain the eigenvalues

$$-3, \quad -\frac{3}{2}\left[1 \pm \sqrt{1 - \frac{8}{3}F_1 Q \frac{d\lambda}{dF}(F_1)}\right], \tag{94}$$

where $F_1 \equiv F(\phi_1)$ is the value of F at the de Sitter point with the field value ϕ_1. Since $F_1 > 0$, we find that the de Sitter point is stable for

$$Q\frac{d\lambda}{dF}(F_1) \geq 0, \quad \text{i.e.,} \quad \frac{d\lambda}{d\phi}(\phi_1) \leq 0. \tag{95}$$

Let us recall the $f(R)$ model (61), which recovers the models (58) and (59) in the regime $R \gg R_c$. Since $e^{2\phi/\sqrt{6}} = 1 - 2n\mu(R/R_c)^{-(2n+1)}$ in this case, the potential $U = (FR - f)/2$ is given by

$$U(\phi) = \frac{\mu R_c}{2}\left[1 - \frac{2n+1}{(2n\mu)^{2n/(2n+1)}}\left(1 - e^{2\phi/\sqrt{6}}\right)^{2n/(2n+1)}\right], \tag{96}$$

in which case the slope of the potential, $\lambda = -U_{,\phi}/U$, is

$$\lambda = -\frac{4ne^{2\phi/\sqrt{6}}}{\sqrt{6}(2n\mu)^{2n/(2n+1)}}\left[1 - \frac{2n+1}{(2n\mu)^{2n/(2n+1)}}\left(1 - e^{2\phi/\sqrt{6}}\right)\right]^{-2n/(2n+1)} \times \left(1 - e^{2\phi/\sqrt{6}}\right)^{-1/(2n+1)}. \tag{97}$$

In the deep matter era during which the condition $R/R_c \gg 1$ is satisfied, the field ϕ is very close to zero. For n and μ of the order of unity, we have $|\lambda| \gg 1$ at this stage. Hence the matter era can be realized by the instantaneous fixed point (d). As R/R_c gets smaller, $|\lambda|$ decreases to the order of unity. If the solutions reach the point $\lambda = 4Q = -4/\sqrt{6}$ and satisfy the stability condition $d\lambda/dF \leq 0$, then the final attractor corresponds to the de Sitter fixed point (e).

For the theories with general couplings Q, it is possible to construct a scalar-field potential that is the generalization of (96). One example is [45]

$$U(\phi) = U_0\left[1 - C(1 - e^{-2Q\phi})^p\right] \qquad (U_0 > 0, C > 0, 0 < p < 1). \tag{98}$$

The $f(R)$ model (61) corresponds to $Q = -1/\sqrt{6}$ and $p = 2n/(2n+1)$. The slope of the potential is given by

$$\lambda = \frac{2Cp\,Qe^{-2Q\phi}(1-e^{-2Q\phi})^{p-1}}{1-C(1-e^{-2Q\phi})^{p}}. \tag{99}$$

We have $U(\phi) \to U_0$ for $\phi \to 0$ and $U(\phi) \to U_0(1-C)$ in the limits $\phi \to \infty$ (for $Q > 0$) and $\phi \to -\infty$ (for $Q < 0$).

The field is nearly frozen around the value $\phi = 0$ during the deep radiation and matter epochs. In these epochs we have $R \simeq \rho_m/F$ from (79), (80), and (81) by noting that U_0 is negligibly small compared to ρ_m or ρ_r. Using (81), it follows that $U_{,\phi} + Q\rho_m \simeq 0$. Hence, in the high-curvature region, the field ϕ evolves along the instantaneous minima given by

$$\phi_m \simeq \frac{1}{2Q}\left(\frac{2U_0pC}{\rho_m}\right)^{\frac{1}{1-p}}. \tag{100}$$

The field value $|\phi_m|$ increases for decreasing ρ_m. As long as the condition $\rho_m \gg 2U_0pC$ is satisfied, we have $|\phi_m| \ll 1$ from (100).

Equation (99) shows that $|\lambda| \gg 1$ for field values around $\phi = 0$. Hence the instantaneous fixed point (d) can be responsible for the matter-dominated epoch provided that $|Q| \ll |\lambda|$. The variable $F = e^{-2Q\phi}$ decreases in time irrespective of the sign of the coupling Q and hence $0 < F < 1$. The de Sitter solution corresponds to $\lambda = 4Q$, that is

$$C = \frac{2}{(1-F_1)^{p-1}\left[2+(p-2)F_1\right]}. \tag{101}$$

The de Sitter solution is present as long as the solution of this equation exists in the region $0 < F_1 < 1$.

From (99) the derivative of λ with respect to ϕ is

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\phi} = -\frac{4CpQ^2F(1-F)^{p-2}[1-pF-C(1-F)^p]}{[1-C(1-F)^p]^2}. \tag{102}$$

The de Sitter point is stable under the condition $1-pF_1 > C(1-F_1)^p$. Using (101) this condition translates into

$$F_1 > 1/(2-p). \tag{103}$$

When $0 < C < 1$, it is possible to show that $\mathrm{d}\lambda/\mathrm{d}\phi < 0$ is always satisfied. Hence the solutions approach the de Sitter attractor after the end of the matter era. When $C > 1$, the de Sitter point is stable under the condition (103). If this condition is violated, the solutions choose another stable fixed point [such as the point (c)] as an attractor.

The above discussion shows that, when $0 < C < 1$, the matter point (d) can be followed by the stable de Sitter solution (e) for the model (98). In Fig. 2 we plot the evolution of $\Omega_{\rm DE}$, Ω_m, $\Omega_{\rm rad}$, and $w_{\rm eff}$ for $Q = 0.01$, $p = 0.2$, and $C = 0.7$.

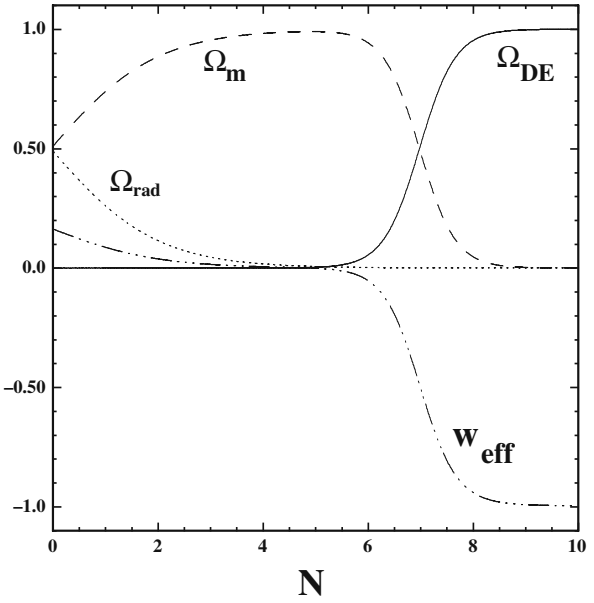

Fig. 2 The evolution of $\Omega_{\rm DE}$, Ω_m, $\Omega_{\rm rad}$, and $w_{\rm eff}$ for the model (98) with parameters $Q = 0.01$, $p = 0.2$, and $C = 0.7$, and initial conditions $x_1 = 0$, $x_2 = 2.27 \times 10^{-7}$, $x_3 = 0.7$, and $x_4 - 1 = -5.0 \times 10^{-13}$

This shows that the viable cosmological trajectory can be realized for the potential (98). In order to confront with SN Ia observations, it is possible to rewrite (79) and (80) in the forms of (30) and (31) by defining the dark energy density $\rho_{\rm DE}$ and the pressure $P_{\rm DE}$ in the similar way. It was shown in [45] that the phantom equation of state as well as the cosmological constant boundary crossing can be realized for the field potentials $U(\phi)$ satisfying local gravity constraints.

3.2 Local Gravity Constraints

We study the local gravity constraints (LGC) for the scalar-tensor theories given by the action (77). In the absence of the potential $U(\phi)$ the Brans–Dicke parameter $\omega_{\rm BD}$ is constrained to be $\omega_{\rm BD} > 4.0 \times 10^4$ from solar system experiments. Note that this bound also applies to the case of a nearly massless field with the potential $U(\phi)$ in which the Yukawa correction e^{-Mr} is close to unity (where M is the scalar-field mass and r is an interaction length). Using the bound $\omega_{\rm BD} > 4.0 \times 10^4$ in (78), we find

$$|Q| < 2.5 \times 10^{-3}. \tag{104}$$

This is a strong constraint under which the cosmological evolution for such theories is difficult to be distinguished from the $Q = 0$ case.

In the presence of the field potential, it is possible for large coupling models ($|Q| = \mathcal{O}(1)$) to satisfy the local gravity constraints provided that the mass M of the field ϕ is sufficiently large in the region of high density. In fact the scalar-tensor potential (98) is designed to have a large mass in the high-density region so that it can be compatible with experimental tests for the violation of equivalence principle through the chameleon mechanism even for $|Q| = \mathcal{O}(1)$. In the following, let us consider the model (98) and derive the conditions under which the local gravity constraints can be satisfied. If we make a conformal transformation for the action (98), the Einstein frame action is given by the action (77) with $F(\phi) = e^{-2Q\phi}$. We can use the results obtained in Sect. 2.2 because we derived thin-shell solutions for the general coupling Q.

As in the case of $f(R)$ gravity, we consider a configuration in which a spherically symmetric body has a constant density ρ_A inside the body and that the energy density outside the body is given by $\rho = \rho_B$ ($\ll \rho_A$). Under the condition $|Q\phi| \ll 1$, we have $V_{,\phi} \simeq -2U_0QpC(2Q\phi)^{p-1}$ for the potential $V = U/F^2$ in the Einstein frame. Then the field values at the potential minima inside and outside the body are

$$\phi_A \simeq \frac{1}{2Q}\left(\frac{2U_0\,p\,C}{\rho_A}\right)^{1/(1-p)}, \quad \phi_B \simeq \frac{1}{2Q}\left(\frac{2U_0\,p\,C}{\rho_B}\right)^{1/(1-p)}. \tag{105}$$

In order to realize the accelerated expansion today, U_0 needs to be roughly the same order as the square of the present Hubble parameter H_0, so we have $U_0 \sim H_0^2 \sim \rho_0$, where $\rho_0 \simeq 10^{-29}$ g/cm^3 is the present cosmological density. The baryonic/dark matter density in our galaxy corresponds to $\rho_B \simeq 10^{-24}$ g/cm^3. Hence the conditions $|Q\phi_A| \ll 1$ and $|Q\phi_B| \ll 1$ are in fact satisfied unless $C \gg 1$. The field mass squared $m_A^2 \equiv V_{,\phi\phi}$ at $\phi = \phi_A$ is approximately given by

$$m_A^2 \simeq \frac{1-p}{(2^p\,pC)^{1/(1-p)}}Q^2\left(\frac{\rho_A}{U_0}\right)^{(2-p)/(1-p)}U_0, \tag{106}$$

which means that m_A can be much larger than H_0 because of the condition $\rho_A \gg U_0$. This large mass allows the chameleon mechanism to work so that the condition $1/(m_A\tilde{r}_c) \ll 1$ is satisfied.

The bound (57) coming from the violation of equivalence principle in solar system translates into

$$(2U_0pC/\rho_B)^{1/(1-p)} < 7.4 \times 10^{-15}\,|Q|. \tag{107}$$

We shall consider the case in which the solutions finally approach the de Sitter point (e). At the de Sitter point (e), one has $3F_1H_1^2 = U_0[1 - C(1 - F_1)^p]$ with C given in (101). Then we get the following relation:

$$U_0 = 3H_1^2\left[2 + (p-2)F_1\right]/p. \tag{108}$$

Plugging this into (107), we find

$$(R_1/\rho_B)^{1/(1-p)}\,(1-F_1) < 7.4 \times 10^{-15}|Q|, \tag{109}$$

where $R_1 = 12H_1^2$ is the Ricci scalar at the de Sitter point. Since the term $(1-F_1)$ is smaller than 1/2 from the condition (103), we obtain the inequality $(R_1/\rho_B)^{1/(1-p)} < 1.5 \times 10^{-14}|Q|$. Using the values $R_1 = 10^{-29}$ g/cm^3 and $\rho_B = 10^{-24}$ g/cm^3, we obtain the following bound:

$$p > 1 - \frac{5}{13.8 - \log_{10}|Q|}. \tag{110}$$

When $|Q| = 10^{-1}$ and $|Q| = 1$, we have $p > 0.66$ and $p > 0.64$, respectively. Thus the model can be compatible with local gravity experiments even for $|Q| = \mathcal{O}(1)$.

4 Braneworld Models of Dark Energy

In this section we discuss braneworld models of dark energy motivated by string theory. In braneworlds standard model particles are confined on a 3D brane embedded in 5D bulk space-time with large extra dimensions. Dvali, Gabadadze, and Porrati (DGP) [23] proposed a braneworld model in which the 3-brane is embedded in a Minkowski bulk space-time with infinitely large extra dimensions. Newton's law can be recovered by adding a 4D Einstein–Hilbert action sourced by the brane curvature to the 5D action [70]. The presence of such a 4D term may be induced by quantum corrections coming from the bulk gravity and its coupling with matter on the brane. In the DGP model the standard 4D gravity is recovered for small distances, whereas the effect from the 5D gravity manifests itself for large distances. Interestingly, it is possible to realize the late-time cosmic acceleration without introducing a dark energy component [71, 72].

The action for the DGP model is given by

$$S = \frac{1}{2\kappa_{(5)}^2}\int \mathrm{d}^5X\sqrt{-\tilde{g}}\,\tilde{R} + \frac{1}{2\kappa_{(4)}^2}\int \mathrm{d}^4X\sqrt{-g}R - \int \mathrm{d}^5X\sqrt{-\tilde{g}}\,\mathcal{L}_m, \tag{111}$$

where $\tilde{g}_{AB}$ is the metric in the 5D bulk and $g_{\mu\nu} = \partial_\mu X^A \partial_\nu X^B \tilde{g}_{AB}$ is the induced metric on the brane with $X^A(x^c)$ being the coordinates of an event on the brane labelled by x^c. The 5D and 4D gravitational constants, $\kappa_{(5)}^2$ and $\kappa_{(4)}^2$, are related wth the 5D and 4D Planck masses, $M_{(5)}$ and $M_{(4)}$, via

$$\kappa_{(5)}^2 = 1/M_{(5)}^3, \qquad \kappa_{(4)}^2 = M_{(4)}^2. \tag{112}$$

The first and second terms in (111) correspond to Einstein–Hilbert actions in the 5D bulk and on the brane, respectively.

There is no contribution to the Lagrangian $\mathcal{L}_m$ from the bulk matter because we are considering a Minkowski bulk. Then the matter action consists of a brane-localized matter whose action is given by $\int \mathrm{d}^4x\sqrt{-g}\,(\sigma + \mathcal{L}_m^{\mathrm{brane}})$, where σ is the 3-brane tension and $\mathcal{L}_m^{\mathrm{brane}}$ is the Lagrangian density on the brane. Since the tension is not related to the Ricci scalar R, it can be adjusted to be zero, as we do in the following.

To study cosmological dynamics on the brane (located at $y = 0$), we take a metric of the form:

$$\mathrm{d}s^2 = -n^2(\tau,y)\mathrm{d}\tau^2 + a^2(\tau,y)\,\gamma_{ij}\mathrm{d}x^i\mathrm{d}x^j + \mathrm{d}y^2, \tag{113}$$

where γ_{ij} represents a maximally symmetric space-time with a constant curvature K. The 5D Einstein equations are given by

$$\tilde{G}_{AB} \equiv \tilde{R}_{AB} - \frac{1}{2}\tilde{R}\,\tilde{g}_{AB} = \kappa_{(5)}^2\tilde{T}_{AB}, \tag{114}$$

where $\tilde{R}_{AB}$ is the 5D Ricci tensor, $\tilde{T}_{AB}$ is the sum of the energy momentum tensor $T_{AB}^{(\mathrm{brane})}$ on the brane and the contribution $\tilde{U}_{AB}$ coming from the scalar curvature of the brane:

$$\tilde{T}_{AB} = T_{AB}^{(\mathrm{brane})} + \tilde{U}_{AB}. \tag{115}$$

Since we are considering a homogeneous and isotropic universe on the brane, one can write $T_B^{A\,(\mathrm{brane})}$ in the form

$$T_B^{A\,(\mathrm{brane})} = \delta(y)\,\mathrm{diag}(-\rho_M,P_M,P_M,P_M,0). \tag{116}$$

Note that ρ_M and P_M are function of τ only. The non-vanishing components coming from the Ricci scalar R of the brane are

$$\tilde{U}_{00} = -\frac{3}{\kappa_{(4)}^2}\left(\frac{\dot{a}^2}{a^2} + K\frac{n^2}{a^2}\right)\delta(y), \tag{117}$$

$$\tilde{U}_{ij} = -\frac{1}{\kappa_{(4)}^2}\left[\frac{a^2}{n^2}\left(-\frac{\dot{a}^2}{a^2} + 2\frac{\dot{a}}{a}\frac{\dot{n}}{n} - 2\frac{\ddot{a}}{a}\right) - K\right]\gamma_{ij}\,\delta(y), \tag{118}$$

where a dot represents a derivative with respect to τ. The non-vanishing components of the 5D Einstein tensor $\tilde{G}_{AB}$ are [73, 74, 71]

$$\tilde{G}_{00} = 3\left[\frac{\dot{a}^2}{a^2} - n^2\left(\frac{a''}{a} + \frac{a'^2}{a^2}\right) + K\frac{n^2}{a^2}\right], \tag{119}$$

$$\tilde{G}_{ij} = \left[a^2\left(2\frac{a''}{a} + \frac{n''}{n} + \frac{a'^2}{a^2} + 2\frac{a'n'}{an}\right) + \frac{a^2}{n^2}\left(-2\frac{\ddot{a}}{a} - \frac{a'^2}{a^2} + 2\frac{\dot{a}\dot{n}}{an}\right) - K\right]\gamma_{ij}, \tag{120}$$

$$\tilde{G}_{05} = 3\left(\frac{\dot{a}n'}{an} - \frac{\dot{a}'}{a}\right), \tag{121}$$

$$\tilde{G}_{55} = 3\left(\frac{a'^2}{a^2} + \frac{a'n'}{an}\right) - \frac{3}{n^2}\left(\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} - \frac{\dot{a}\dot{n}}{an}\right) - 3\frac{K}{a^2}, \tag{122}$$

where a prime represents a derivative with respect to y.

Assuming no flow of matter along the fifth dimension, we have $\tilde{T}_{05} = 0$ and hence $\tilde{G}_{05} = 0$. We then find that (119) and (122) can be written as

$$\tilde{G}_{00} = -\frac{3n^2}{2a^3a'}I', \qquad \tilde{G}_{55} = -\frac{3}{2a^3\dot{a}}\dot{I}, \tag{123}$$

where

$$I \equiv (a'a)^2 - \frac{(\dot{a}a)^2}{n^2} - Ka^2. \tag{124}$$

Since we are considering the Minkowski bulk, we have $\tilde{G}_{00} = 0$ and $\tilde{G}_{55} = 0$ locally in the bulk. This then gives $I' = 0$ and $\dot{I} = 0$. The integration of these equations leads to

$$(a'a)^2 - \frac{(\dot{a}a)^2}{n^2} - Ka^2 + C = 0, \tag{125}$$

where C is a constant independent of τ and y.

Let us find solutions of the Einstein equations (114) in the vicinity of $y = 0$. The metric needs to be continuous across the brane in order to have a well-defined geometry. Note, however, that its derivatives with respect to y can be discontinuous at $y = 0$. The Einstein tensor is made of the metric up to the second derivatives with respect to y, so the Einstein equations with a distributional source are written in the form [71, 73, 74]

$$g'' = T\,\delta(y), \tag{126}$$

where $\delta(y)$ is a Dirac's delta function. Integrating this equation across the brane gives

$$[g'] = T, \quad \text{where} \quad [g'] \equiv g'(0^+) - g'(0^-). \tag{127}$$

The jump of the first derivative of the metric is equivalent to the energy momentum tensor on the brane.

Equations (119) and (120) include the second derivatives a'' and n'' of the metric. Integrating the Einstein equations, $\tilde{G}_{00} = \kappa_{(5)}^2 \tilde{T}_{00}$ and $\tilde{G}_{ij} = \kappa_{(5)}^2 \tilde{T}_{ij}$, across the brane, we obtain

$$\frac{[a']}{a_b} = -\frac{\kappa_{(5)}^2}{3}\rho_M + \frac{\kappa_{(5)}^2}{\kappa_{(4)}^2 n_b^2}\left(\frac{\dot{a}_b^2}{a_b^2} + K\frac{n_b^2}{a_b^2}\right), \tag{128}$$

$$\frac{[n']}{n_b} = \frac{\kappa_{(5)}^2}{3}(3P_M + 2\rho_M) - \frac{\kappa_{(5)}^2}{\kappa_{(4)}^2 n_b^2}\left(\frac{\dot{a}_b^2}{a_b^2} + 2\frac{\dot{a}_b}{a_b}\frac{\dot{n}_b}{n_b} - 2\frac{\ddot{a}_b}{a_b} + K\frac{n_b^2}{a_b^2}\right), \tag{129}$$

where the subscript "b" represents the quantities on the brane.

We assume the symmetry $y \leftrightarrow -y$, in which case $[a'] = 2a'(0^+)$ and $[n'] = 2n'(0^+)$. Substituting (128) into (125), we obtain the modified Friedmann equation on the brane:

$$\varepsilon\sqrt{H^2 + \frac{K}{a_b^2} - \frac{C}{a_b^4}} = \frac{\kappa_{(5)}^2}{2\kappa_{(4)}^2}\left(H^2 + \frac{K}{a_b^2}\right) - \frac{\kappa_{(5)}^2}{6}\rho_M, \tag{130}$$

where $H \equiv \dot{a}_b/(a_b n_b)$ is the Hubble parameter and $\varepsilon = \pm 1$ is the sign of $[a']$. The constant C can be interpreted as the term coming from the 5D bulk Weyl tensor [71, 72, 75]. Since the Weyl tensor vanishes for the Minkowski bulk, we set $C = 0$ in the following discussion. We also introduce a length scale

$$r_c \equiv \frac{\kappa_{(5)}^2}{2\kappa_{(4)}^2} = \frac{M_{(4)}^2}{2M_{(5)}^3}. \tag{131}$$

Then (130) can be written as

$$\frac{\varepsilon}{r_c}\sqrt{H^2 + \frac{K}{a^2}} = H^2 + \frac{K}{a^2} - \frac{\kappa_{(4)}^2}{3}\rho_M, \tag{132}$$

where we have omitted the subscript "b" for the quantities at $y = 0$.

Plugging the junction conditions (128) and (129) into the (05) component of the Einstein equations, $\tilde{G}_{05} = 0$, the following matter conservation equation holds on the brane:

$$\frac{\mathrm{d}\rho_M}{\mathrm{d}t} + 3H(\rho_M + P_M) = 0, \tag{133}$$

where t is the cosmic time related to the time τ via the relation $\mathrm{d}t = n_b \mathrm{d}\tau$. If the equation of state, $w_M = P_M/\rho_M$, is specified, the cosmological evolution is obtained by solving (132) and (133).

For a flat geometry ($K = 0$), (132) reduces to

$$H^2 - \frac{\varepsilon}{r_c}H = \frac{\kappa_{(4)}^2}{3}\rho_M. \tag{134}$$

If the crossover scale r_c is much larger than the Hubble radius H^{-1}, the first term in (134) dominates over the second one. In this case the standard Friedmann equation, $H^2 = \kappa_{(4)}^2\rho_M/3$, is recovered. Meanwhile, in the regime $r_c < H^{-1}$, the presence of the second term in (134) leads to a modification to the standard Friedmann equation. In the universe dominated by non-relativistic matter ($\rho_M \propto a^{-3}$), the universe approaches a de Sitter solution for $\varepsilon = +1$:

$$H \to H_{\rm dS} = \frac{1}{r_c}. \tag{135}$$

Hence it is possible to realize the present cosmic acceleration provided that r_c is of the order of the present Hubble radius H_0^{-1}.

Equation (132) can be written as

$$H^2 + \frac{K}{a^2} = \left(\sqrt{\frac{\kappa_{(4)}^2}{3}\rho_M + \frac{1}{4r_c^2}} + \frac{1}{2r_c}\right)^2. \tag{136}$$

For the matter on the brane, we consider non-relativistic matter with the energy density ρ_m and the equation of state $w_m = 0$. We then have $\rho_m = \rho_m^{(0)}(1+z)^3$ from (133). We introduce the following present value quantities:

$$\Omega_K^{(0)} = -\frac{K}{a_0^2H_0^2}, \quad \Omega_{r_c}^{(0)} = \frac{1}{4r_c^2H_0^2}, \quad \Omega_m^{(0)} = \frac{\kappa_{(4)}^2\rho_m^{(0)}}{3H_0^2}. \tag{137}$$

Then (136) reads

$$H^2(z) = H_0^2\left[\Omega_K^{(0)}(1+z)^2 + \left\{\sqrt{\Omega_m^{(0)}(1+z)^3 + \Omega_{r_c}^{(0)}} + \sqrt{\Omega_{r_c}^{(0)}}\right\}^2\right]. \tag{138}$$

The normalization condition at $z = 0$ is given by

$$\Omega_m^{(0)} + \Omega_K^{(0)} + 2\sqrt{1 - \Omega_K^{(0)}}\sqrt{\Omega_{r_c}^{(0)}} = 1. \tag{139}$$

For the flat universe ($K = 0$) this relation corresponds to

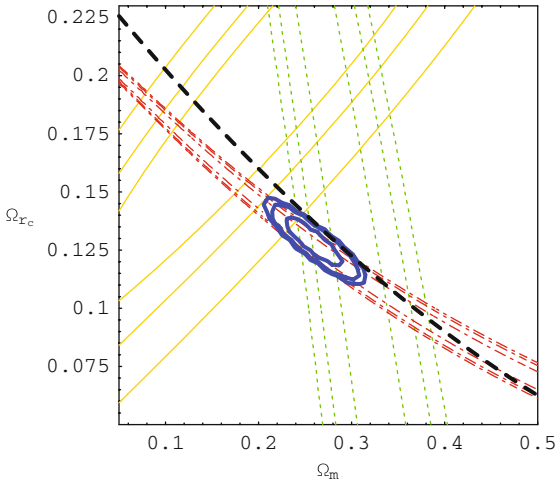

Fig. 3 Observational constraints on the DGP model from the SNLS data (*solid thin*), the BAO (*dotted*), and the CMB shift parameter from the WMAP 3-year data (*dot-dashed*). The *thick line* represents the curve (140) for the flat model ($\Omega_K^{(0)} = 0$). The figure labels Ω_m and Ω_{r_c} correspond to $\Omega_m^{(0)}$ and $\Omega_{r_c}^{(0)}$, respectively. From [52]

$$\Omega_{r_c}^{(0)} = \left(\frac{1 - \Omega_m^{(0)}}{2} \right)^2 . \tag{140}$$

The parametrization (138) of the Hubble parameter together with the normalization (139) can be used to place observational constraints on the DGP model at the background level [51, 52, 54]. In [50], the authors found a significantly worse fit to supernova data and the distance to the last-scattering surface in the pure DGP model as compared to the ΛCDM model. A similar conclusion was reached in [51] and [53], where the two groups have also tried to constrain the DGP model using SN Ia data and the baryon acoustic peak in the Sloan Digital Sky Survey. In Fig. 3 the joint constraints from observational data of SNLS, BAO, and the CMB shift parameter are plotted [52]. While the flat DGP model can be consistent with the SN Ia data, it is under strong observational pressure by adding the data of the BAO and the CMB shift parameter. The open DGP model gives a slightly better fit relative to the flat model [52, 54].

As we will see in the next section, the analysis of 5D cosmological perturbations shows that the DGP model contains a ghost mode in the scalar sector of the gravitational field [55–57]. Hence the original DGP model is effectively ruled out as a viable dark energy model by the observational pressure and by the ghost problem.

5 Other Modified Gravity Models

There are other classes of modified gravity models in which the Lagrangian density f is an arbitrary function of R, $P \equiv R_{\mu\nu}R^{\mu\nu}$ and $Q \equiv R_{\mu\nu\alpha\beta}R^{\mu\nu\alpha\beta}$, where $R_{\mu\nu}$ and $R_{\mu\nu\alpha\beta}$ are Ricci tensor and Riemann tensor, respectively [76]. In order to avoid

the appearance of spurious spin-2 ghosts, we need to take a Gauss–Bonnet (GB) combination [77–80], i.e.,

$$R^2_{\rm GB} = R^2 - 4R_{\mu\nu}R^{\mu\nu} + R_{\mu\nu\alpha\beta}R^{\mu\nu\alpha\beta}. \tag{141}$$

The simple model that can give rise to cosmic acceleration is provided by the action

$$S = \int d^4x\sqrt{-g}\left[\frac{1}{2}R - \frac{1}{2}(\nabla\phi)^2 - V(\phi) - f(\phi)R^2_{\rm GB}\right] + S_m, \tag{142}$$

where $V(\phi)$ and $f(\phi)$ are the functions of a scalar field ϕ and S_m is a matter action. For the exponential potential $V(\phi) = V_0 e^{-\lambda\phi}$ and the coupling $f(\phi) = (f_0/\mu)e^{\mu\phi}$, the cosmological dynamics has been extensively studied in [24, 81–84]. In particular, it was found in [81, 82] that a scaling matter era can be followed by a late-time de Sitter solution that appears due to the presence of the GB term.

Koivisto and Mota [81] placed observational constraints on the above model using the Gold data set of SN Ia together with the CMB shift parameter data of WMAP. The parameter λ is constrained to be $3.5 < \lambda < 4.5$ at the 95% confidence level. In the second paper [83], they included the constraints coming from the BBN, LSS, BAO, and solar system data and showed that these data strongly disfavor the GB model discussed above. Moreover, it was shown in [82] that tensor perturbations are subject to negative instabilities in the above model when the GB term dominates the dynamics (see also [85]). The paper [86] studied local gravity constraints on the GB models with couplings of the form $f(\phi)R^2_{\rm GB}$ and showed that the energy contribution coming from the GB term needs to be strongly suppressed to be compatible with solar system experiments. This is typically of the order of $\Omega_{\rm GB} < 10^{-30}$ and hence the GB term of the coupling $f(\phi)R^2_{\rm GB}$ cannot be responsible for the current accelerated expansion of the universe. The above discussions show that the GB term with the scalar-field coupling $f(\phi)R^2_{\rm GB}$ can hardly be the source for dark energy.

The models based on the Lagrangian density $\mathcal{L} = R/2 + f(\mathcal{G})$, where $\mathcal{G} = R^2_{\rm GB}$ is the GB term, have been studied by a number of authors [87–89]. In order to ensure the stability of a late-time de Sitter solution and radiation/matter solutions, we need to satisfy the condition $\partial^2 f/\partial\mathcal{G}^2 > 0$ [89]. In [89] the authors presented a number of $f(\mathcal{G})$ models that are cosmologically viable at least at the background level (see also [90]). One of such viable models is given by

$$f(\mathcal{G}) = \lambda\frac{\mathcal{G}}{\sqrt{\mathcal{G}_*}}\arctan\left(\frac{\mathcal{G}}{\mathcal{G}_*}\right) - \alpha\lambda\sqrt{\mathcal{G}_*}, \tag{143}$$

where α, λ, and $\mathcal{G}_*$ are constants. It will be of interesting to study whether such models can be consistent with local gravity tests.

6 Observational Signatures of Modified Gravity

In order to confront modified gravity models with observations such as large-scale structure and CMB, we discuss the evolution of density perturbations in three modified gravity models: (i) $f(R)$ gravity, (ii) scalar-tensor gravity, and (iii) DGP braneworld model. This is important to distinguish modified gravity models from the ΛCDM model. We also discuss observables to confront with weak lensing observations.

6.1 f(R) Gravity

Let us consider the case of $f(R)$ gravity in the metric formalism in the presence of non-relativistic matter. We adopt the following perturbed metric with scalar metric perturbations Φ and Ψ in a longitudinal gauge about the flat FLRW background

$$ds^2 = a^2[-(1+2\Psi)d\eta^2 + (1+2\Phi)\delta_{ij}\,dx^i\,dx^j], \tag{144}$$

where $\eta = \int a^{-1}dt$ is the conformal time. The energy momentum tensors of non-relativistic matter are decomposed into background and perturbed parts as $T^0_0 = -(\rho_m + \delta\rho_m)$ and $T^0_\alpha = -\rho_m v_{m,\alpha}$, where v_m is a velocity potential.

The Fourier-transformed perturbation equations for matter perturbations are given by [91, 92]

$$\delta\dot{\rho}_m + 3H\delta\rho_m = -\rho_m\left(3\dot{\Phi} + \frac{k^2}{a}v_m\right), \tag{145}$$

$$\dot{v}_m + Hv_m = \frac{1}{a}\Psi, \tag{146}$$

where k is a comoving wavenumber. We define the gauge-invariant matter density perturbation δ_m as

$$\delta_m \equiv \frac{\delta\rho_m}{\rho_m} + 3Hv, \quad \text{where} \quad v \equiv av_m. \tag{147}$$

Then (145) and (146) yield

$$\dot{\delta}_m = -\frac{k^2}{a^2}v - 3(\Phi - Hv)^{\cdot}, \tag{148}$$

$$\dot{v} = \Psi, \tag{149}$$

from which we obtain

$$\ddot{\delta}_m + 2H\dot{\delta}_m + \frac{k^2}{a^2}\Psi = 3\ddot{B} + 6H\dot{B}, \tag{150}$$

where $B \equiv -\Phi + Hv$. For the modes deep inside the Hubble radius ($k \gg aH$) the r.h.s. of (150) can be neglected relative to the l.h.s., so that we approximately obtain

$$\ddot{\delta}_m + 2H\dot{\delta}_m + \frac{k^2}{a^2}\Psi = 0. \tag{151}$$

In $f(R)$ gravity the quantity $F(R) = \partial f/\partial R$ has a perturbation δF. In the following we use the unit $\kappa^2 = 8\pi G = 1$, but we restore gravitational constant G when it is required. Perturbing (2), we obtain the following equations [92]:

$$\begin{aligned} &-\frac{k^2}{a^2}\Phi + 3H(H\Psi - \dot{\Phi}) \\ &= \frac{1}{2F}\left[3H\dot{\delta F} - \left(3\dot{H} + 3H^2 - \frac{k^2}{a^2}\right)\delta F - 3H\dot{F}\Psi - 3\dot{F}(H\Psi - \dot{\Phi}) - \delta\rho_m\right], \end{aligned} \tag{152}$$

$$\begin{aligned} &\ddot{\delta F} + 3H\dot{\delta F} + \left(\frac{k^2}{a^2} - \frac{R}{3}\right)\delta F \\ &= \frac{1}{3}\delta\rho_m + \dot{F}(3H\Psi + \dot{\Psi} - 3\dot{\Phi}) + (2\ddot{F} + 3H\dot{F})\Psi - \frac{1}{3}F\delta R, \end{aligned} \tag{153}$$

$$\Psi + \Phi = -\frac{\delta F}{F}. \tag{154}$$

For the sub-horizon modes satisfying $k \gg aH$, the terms including k^2/a^2 and $\delta\rho_m$ in (152) are the dominant contributions. We then obtain the following approximate relations from (152) and (154) :

$$\Phi = \frac{1}{2F}\left(\frac{a^2}{k^2}\delta\rho_m - \delta F\right), \quad \Psi = -\frac{1}{2F}\left(\frac{a^2}{k^2}\delta\rho_m + \delta F\right). \tag{155}$$

As long as the conditions $|\dot{F}| < |HF|$ and $|\ddot{F}| < |H^2F|$ are satisfied, the second and third terms on the r.h.s. of (153) are much smaller than $\delta\rho_m$ and $(k^2/a^2)\delta F$ for the modes deep inside the Hubble radius. Using the relation $\delta R = \delta F/f_{,RR}$, we find that (153) is approximately given by

$$\ddot{\delta F} + 3H\dot{\delta F} + \left(\frac{k^2}{a^2} + M^2\right)\delta F = \frac{1}{3}\delta\rho_m, \tag{156}$$

where

$$M^2 \equiv \frac{f_{,R}}{3f_{,RR}}. \tag{157}$$

In order to derive (156), we have used the following condition:

$$\left\{\frac{k^2}{a^2}, M^2\right\} \gg R \sim H^2. \tag{158}$$

The condition $M^2 \gg R$ is satisfied for viable $f(R)$ models in the past cosmic expansion history of the universe [40, 42]. We recall that the conditions $f_{,R} > 0$ and $f_{,RR} > 0$ need to be satisfied for $R > R_1$ for the consistency with cosmological and local gravity constraints, so that the mass squared M^2 is positive. In the following we shall discuss two cases: (A) $M^2 \gg k^2/a^2$ and (B) $M^2 \ll k^2/a^2$, separately. For viable $f(R)$ models the mass squared M^2 is large in the past and gradually decreases with time. Hence the transition from the region (A) to the region (B) can occur in the past, depending on the modes k. In the following we consider the model (61) that corresponds to the asymptotic form of the models (58) and (59) in the region $R \gg R_c$.

6.1.1 Evolution of Perturbations in the Regime: $M^2 \gg k^2/a^2$

The solutions for (156) are given by the sum of the oscillating solution $\delta F_{\rm osc}$ obtained by setting $\delta\rho_m = 0$ and the special solution $\delta F_{\rm ind}$ of (156) induced by the presence of matter perturbations $\delta\rho_m$. The oscillating part $\delta F_{\rm osc}$ satisfies the equation $(a^{3/2}\delta F_{\rm osc})^{\cdot\cdot} + M^2(a^{3/2}\delta F_{\rm osc}) \simeq 0$. Using the WKB approximation, the solution is given by [40]

$$\delta F_{\rm osc} \propto a^{-3/2} f_{,RR}^{1/4} \cos\left(\int \frac{1}{\sqrt{3f_{,RR}}}{\rm d}t\right), \tag{159}$$

where we have used $f_{,R} \simeq 1$ because the viable $f(R)$ models are close to the ΛCDM model in the region of high density.

During the matter era in which the background Ricci scalar evolves as $R^{(0)} = 4/(3t^2)$, the quantity $f_{,RR}$ has a dependence $f_{,RR} \propto R^{-2(n+1)} \propto t^{4(n+1)}$ for the model (61). Then the evolution of the perturbation, $\delta R_{\rm osc} = \delta F_{\rm osc}/f_{,RR}$, is given by

$$\delta R_{\rm osc} \simeq c\, t^{-(3n+4)} \cos\left(c_0\, t^{-2(n+1)}\right), \tag{160}$$

where c and c_0 are constants. Unless the coefficient c is chosen to be very small, the perturbation $\delta R_{\rm osc}$ dominates over $R^{(0)}(\propto t^{-2})$ as we go back to the past. This leads to the violation of the stability conditions ($f_{,RR} > 0$ and $f_{,R} > 0$) because the Ricci scalar can be negative.

The special solution $\delta F_{\rm ind}$ for (156) can be derived by neglecting the first and second terms relative to others, giving

$$\delta F_{\rm ind} \simeq \frac{\delta\rho_m}{3M^2}, \qquad \delta R_{\rm ind} \simeq \delta\rho_m. \tag{161}$$

Under the condition $|\delta F_{\rm osc}| \ll |\delta F_{\rm ind}|$, we have $\delta F \simeq \delta\rho_m/(3M^2)$ so that (155) reduces to

$$\Psi = -\Phi = -\frac{1}{2F}\frac{a^2}{k^2}\delta\rho_m. \tag{162}$$

Plugging (162) into (151), we find that the matter perturbation obeys the following equation:

$$\ddot{\delta_m} + 2H\dot{\delta_m} - 4\pi G\rho_m\delta_m/F = 0, \tag{163}$$

where we have reproduced the gravitational constant G for clarity. During the matter-dominated era ($\Omega_m = \rho_m/(3FH^2) = 1$), this has the growing-mode solution

$$\delta_m \propto t^{2/3}. \tag{164}$$

From (161) we get

$$\delta F_{\rm ind} \propto t^{4(n+2/3)}, \qquad \delta R_{\rm ind} \propto t^{-4/3}. \tag{165}$$

Compared to the oscillating mode (160), the matter-induced mode $\delta R_{\rm ind}$ decreases more slowly and thus dominates at late times. Relative to the background value $R^{(0)}$, the perturbation $\delta R = \delta R_{\rm osc} + \delta R_{\rm ind}$ evolves as

$$\frac{\delta R}{R^{(0)}} \simeq c_1\, t^{-(3n+2)} \cos\left(c_0 t^{-p}\right) + c_2\, t^{2/3}, \tag{166}$$

where c_1 and c_2 are constants. In order to avoid the dominance of the oscillating mode at the early epoch, the coefficient c_1 needs to be suppressed relative to c_2 [40, 42].

6.1.2 Evolution of Perturbations in the Regime: $M^2 \ll k^2/a^2$

Since the scalaron mass decreases as $M \propto t^{-2(n+1)}$, the modes that initially exist in the region $M^2 \gg k^2/a^2$ can enter the regime $M^2 \ll k^2/a^2$ during the matter-dominated epoch. It is sufficient to consider the matter-induced mode because the oscillating mode is already suppressed during the evolution in the regime $M^2 \gg k^2/a^2$. The matter-induced special solution of (156) in the regime $M^2 \ll k^2/a^2$ is approximately given by

$$\delta F_{\rm ind} \simeq \frac{a^2}{3k^2}\delta\rho_m. \tag{167}$$

From (155) the gravitational potentials satisfy

$$\Psi = -\frac{4}{3}\cdot\frac{1}{2F}\frac{a^2}{k^2}\delta\rho_m, \quad \Phi = \frac{2}{3}\cdot\frac{1}{2F}\frac{a^2}{k^2}\delta\rho_m. \tag{168}$$

Plugging (168) into (151), the matter perturbation obeys the following equation:

$$\ddot{\delta_m} + 2H\dot{\delta_m} - \frac{4}{3}\cdot 4\pi G\rho_m\delta_m/F = 0. \tag{169}$$

During the matter-dominated epoch with $\Omega_m \simeq 1$ and $a \propto t^{2/3}$, we obtain the following evolution:

$$\delta_m \propto t^{\frac{\sqrt{33}-1}{6}}. \tag{170}$$

The growth rate of δ_m gets larger compared to (164).

6.1.3 Matter Power Spectra

If the transition from the regime $M^2 \gg k^2/a^2$ to the regime $M^2 \ll k^2/a^2$ occurs during the matter era, the evolution of matter perturbations changes from $\delta_m \propto t^{2/3}$ to $\delta_m \propto t^{(\sqrt{33}-1)/6}$. We use the subscript "$k$" for the quantities at which k is equal to aM, whereas the subscript "Λ" is used at which the accelerated expansion starts ($\ddot{a} = 0$). While the redshift z_Λ is independent of k, z_k depend on k and also on the mass M.

For the model (61) the variable $m = Rf_{,RR}/f_{,R}$ can grow fast from the regime $m \ll (aH/k)^2$ (i.e., $M^2 \gg k^2/a^2$) to the regime $m \gg (aH/k)^2$ (i.e., $M^2 \ll k^2/a^2$). In fact, m can grow to as large as the order of 0.1 even if m is much smaller than 10^{-6} in the deep matter era. For the sub-horizon modes relevant to the galaxy power spectrum, the transition at $M^2 = k^2/a^2$ typically occurs at the redshift z_k larger than 1 (provided that $n = \mathcal{O}(1)$). For the mode $k/(a_0H_0) = 300$, one has $z_k = 4.83$ for $n = 1$ and $z_k = 2.49$ for $n = 2$. As n gets larger, the period of a non-standard evolution of δ_m becomes shorter because z_k tends to be smaller. Since the scalaron mass evolves as $M \propto t^{-2(n+1)}$ for the model (61), the time t_k has a scale dependence $t_k \propto k^{-\frac{3}{6n+4}}$. This means that the smaller-scale modes cross the transition point earlier. The matter power spectrum $P_{\delta_m} = |\delta_m|^2$ at the time t_Λ shows a difference compared to the case of the ΛCDM model:

$$\frac{P_{\delta_m}(t_\Lambda)}{P_{\delta_m}^{\Lambda\mathrm{CDM}}(t_\Lambda)} = \left(\frac{t_\Lambda}{t_k}\right)^{2\left(\frac{\sqrt{33}-1}{6}-\frac{2}{3}\right)} \propto k^{\frac{\sqrt{33}-5}{6n+4}}. \tag{171}$$

The galaxy matter power spectrum is modified by this effect. Meanwhile the CMB spectrum is hardly affected except for very large scales (for the multipoles $\ell = \mathcal{O}(1)$) at which the integrated Sachs–Wolfe (ISW) effect becomes important. Hence there is a difference for the spectral indices of two power spectra, i.e.,

$$\Delta n(t_\Lambda) = \frac{\sqrt{33}-5}{6n+4}. \tag{172}$$

For larger n the redshift z_k can be as close as z_Λ, which means that the estimation (172) is not necessarily valid in such cases. Moreover, the estimation (172) does not take into account the evolution of δ_m after $z = z_\Lambda$ to the present epoch ($z = 0$). It was found in [42] that the estimation (172) agrees well with the numerically obtained $\Delta n(t_\Lambda)$ for $n \leq 2$.

After the system enters the epoch of cosmic acceleration, the momentum k can again become smaller than aM. Hence the k-dependence is not necessarily negligible even for $z < z_\Lambda$. However, we find that $\Delta n(t_0)$ is not much different from $\Delta n(t_\Lambda)$ derived by (172). Thus the analytic estimation (172) is certainly reliable to place constraints on model parameters except for $n \gg 1$. Observationally, we do not find any strong difference for the slopes of the spectra of LSS and CMB. If we take the mild bound $\Delta n(t_\Lambda) < 0.05$, we obtain the constraint $n \geq 2$. In this case the local gravity constraint (67) is also satisfied.

The modified growth of matter perturbations also affects the evolution of the gravitational potentials Ψ and Φ. The effective potential $\psi \equiv \Phi - \Psi$ is important to discuss the ISW effect on the CMB as well as the weak lensing observations [93]. From (155) this potential is given by

$$\psi = \frac{3a^2H^2}{k^2}\Omega_m\delta_m. \tag{173}$$

In the ΛCDM model the potential ψ remains constant during the standard matter era, but it decays after the system enters the accelerated epoch, producing the ISW contribution for low multipoles on the CMB power spectrum. In $f(R)$ gravity the additional growth of matter perturbations in the region $z < z_k$ changes the evolution of ψ.

From CMB observations, however, we do not obtain a constraint on n tighter than the one derived by the spectral index of matter perturbations [94]. This comes from the fact that the ISW effect is important only for the modes with $k/(a_0H_0) = \mathcal{O}(1)$ whose transition redshift z_k is smaller than the modes relevant to the galaxy power spectrum. In the weak lensing observations, the modified evolution of the lensing potential ψ directly leads to the change even for the small-scale shear power spectrum [95, 96]. Hence this can be a powerful tool to constrain $f(R)$ gravity models from future observations.

6.2 Scalar-Tensor Gravity

We shall next discuss the case of scalar-tensor gravity. To be concrete we shall study the evolution of matter perturbations for the Jordan frame action (77), i.e., Brans–Dicke theory with the potential $U(\phi)$ and the coupling $F(\phi) = e^{-2Q\phi}$. We define the field mass squared to be

$$M^2 \equiv U_{,\phi\phi}. \tag{174}$$

If the scalar field is light such that the condition $M < H_0$ is always satisfied irrespective of high- or low-density regions, the coupling Q is constrained to be $|Q| < 10^{-3}$ from local gravity tests. Meanwhile, if the mass M in the region of high density is much larger than that on cosmological scales, it is possible to satisfy local gravity constraints by the chameleon mechanism even if $|Q|$ is of the order of unity. Cosmologically the mass M can decrease from the past to the present, which can allow the transition from the "GR regime" to the "scalar-tensor regime" as it happens in $f(R)$ gravity. An example of field potential showing this behavior is given by (98).

As in $f(R)$ gravity, the matter perturbation δ_m satisfies (151). The difference appears in the expression of the gravitational potential Ψ. In Fourier space the scalar metric perturbations obey the following equations [92, 45]:

$$-\frac{k^2}{a^2}\Phi + 3H(H\Psi - \dot{\Phi}) = -\frac{1}{2F}\Bigg[\omega\dot{\phi}\delta\dot{\phi} + \frac{1}{2}(\omega_{,\phi}\dot{\phi}^2 - F_{,\phi}R + 2V_{,\phi})\delta\phi + \left(3\dot{H} + 3H^2 - \frac{k^2}{a^2}\right)\delta F - 3H\delta\dot{F} + (3H\dot{F} - \omega\dot{\phi}^2)\Psi + 3\dot{F}(H\Psi - \dot{\Phi}) + \delta\rho_m\Bigg], \tag{175}$$

$$\delta\ddot{\phi} + \left(3H + \frac{\omega_{,\phi}}{\omega}\dot{\phi}\right)\delta\dot{\phi} + \left[\frac{k^2}{a^2} + \left(\frac{\omega_{,\phi}}{\omega}\right)_{,\phi}\frac{\dot{\phi}^2}{2} + \left(\frac{2U_{,\phi} - F_{,\phi}R}{2\omega}\right)_{,\phi}\right]\delta\phi = \dot{\phi}\dot{\Psi} + \left(2\ddot{\phi} + 3H\dot{\phi} + \frac{\omega_{,\phi}}{\omega}\dot{\phi}^2\right)\Psi + 3\dot{\phi}(H\Psi - \dot{\Phi}) + \frac{1}{2\omega}F_{,\phi}\delta R, \tag{176}$$

$$\Psi + \Phi = -\frac{\delta F}{F} = -\frac{F_{,\phi}}{F}\delta\phi, \tag{177}$$

where $\delta\phi$ is the perturbed field, $\omega = (1 - 6Q^2)F$ and

$$\delta R = 2\left[3(\dot{\Phi} - H\Psi)^{\cdot} - 12H(H\Psi - \dot{\Phi}) + \left(\frac{k^2}{a^2} - 3\dot{H}\right)\Psi + 2\frac{k^2}{a^2}\Phi\right]. \tag{178}$$

Provided that the mass M defined in (174) is sufficiently heavy to satisfy the conditions $M^2 \gg R$, we can approximate $[(2U_{,\phi} - F_{,\phi}R)/2\omega]_{,\phi} \simeq M^2/\omega$ in (176). The solution for (176) consists of the sum of the matter-induced mode $\delta\phi_{\rm ind}$ sourced by the matter perturbation and the oscillating mode $\delta\phi_{\rm osc}$, i.e., $\delta\phi = \delta\phi_{\rm ind} + \delta\phi_{\rm osc}$ (as in the case of $f(R)$ gravity).

We first derive the matter-induced mode on sub-horizon scales. We use the approximation that the terms containing k^2/a^2, $\delta\rho_m$, δR, and M^2 are the dominant contributions in (175), (176), (177), and (178). Note that this approximation was first used in [65] for the scalar-tensor theory with the Lagrangian density $\mathcal{L} = (1/2)F(\phi)R - (1/2)\nabla(\phi)^2 - U(\phi)$ in the massless limit: $M^2 \ll k^2/a^2$. Under this approximation, we have $\delta R_{\rm ind} \simeq 2(k^2/a^2)[\Phi - (F_{,\phi}/F)\delta\phi_{\rm ind}]$ from (175) and (178), where the subscript "ind" represents the matter-induced mode. Then from (176) we find

$$\delta\phi_{\rm ind} \simeq -\frac{2QF}{(k^2/a^2)(1-2Q^2)F+M^2}\frac{k^2}{a^2}\Phi. \tag{179}$$

Using (175) and (177) we obtain

$$\frac{k^2}{a^2}\Psi \simeq -\frac{\delta\rho_m}{2F}\frac{(k^2/a^2)(1+2Q^2)F+M^2}{(k^2/a^2)F+M^2}, \tag{180}$$

$$\frac{k^2}{a^2}\Phi \simeq \frac{\delta\rho_m}{2F}\frac{(k^2/a^2)(1-2Q^2)F+M^2}{(k^2/a^2)F+M^2}. \tag{181}$$

In the massive limit $M^2/F \gg k^2/a^2$, we recover the standard result of General Relativity. In the massless limit $M^2/F \ll k^2/a^2$, one has $(k^2/a^2)\Psi \simeq -(\delta\rho_m/2F)(1+2Q^2)$ and $(k^2/a^2)\Phi \simeq (\delta\rho_m/2F)(1-2Q^2)$. Note that this recovers (168) in $f(R)$ gravity by setting $Q=-1/\sqrt{6}$.

Plugging (181) into (151), we obtain the equation for matter perturbations [45]

$$\ddot{\delta}_m + 2H\dot{\delta}_m - 4\pi G_{\rm eff}\rho_m\delta_m = 0, \tag{182}$$

where the effective (cosmological) gravitational "constant" is

$$G_{\rm eff} = \frac{G}{F}\frac{(k^2/a^2)(1+2Q^2)F+M^2}{(k^2/a^2)F+M^2}. \tag{183}$$

We have recovered the bare gravitational constant G. In the massless limit this reduces to

$$G_{\rm eff} \simeq \frac{G}{F}(1+2Q^2) = \frac{G}{F}\frac{4+2\omega_{\rm BD}}{3+2\omega_{\rm BD}} \qquad (M^2 \ll k^2/a^2), \tag{184}$$

where in the last line we have used the relation (78) between the coupling Q and the Brans–Dicke parameter $\omega_{\rm BD}$. In $f(R)$ gravity we have $\omega_{\rm BD}=0$ and hence $G_{\rm eff} = 4G/(3F)$.

Let us derive the approximate equation for the oscillating mode. Using (175) and (176) under the condition $k^2/a^2 \gg H^2$, the gravitational potentials for $\delta\rho_m = 0$ are expressed by $\varphi_{\rm osc}$. Then from (178) the perturbation δR corresponding to the oscillating mode is given by

$$\delta R_{\rm osc} \simeq 6Q\left(\delta\ddot{\phi}_{\rm osc} + 3H\delta\dot{\phi}_{\rm osc} + \frac{k^2}{a^2}\delta\phi_{\rm osc}\right). \tag{185}$$

Substituting this relation in (176), we find

$$\delta\ddot{\phi}_{\rm osc} + 3H\delta\dot{\phi}_{\rm osc} + \left(\frac{k^2}{a^2}+\frac{M^2}{F}\right)\delta\phi_{\rm osc} \simeq 0, \tag{186}$$

which is valid in the regime $M^2 \gg R$.

When $|Q| = \mathcal{O}(1)$ the field potential $U(\phi)$ is required to be heavy in the region of high density for the consistency with local gravity constraints. We shall consider the potential (98) as an example of a viable model. During the matter era the field ϕ sits at the instantaneous minima characterized by the condition (100). Then we have the relations $\phi \propto \rho_m^{\frac{1}{p-1}}$ and $M^2 \propto \rho_m^{\frac{2-p}{1-p}}$ during the matter-dominated epoch. The field ϕ can initially be heavy to satisfy the condition $M^2/F \gg k^2/a^2$ for the modes relevant to the galaxy power spectrum. Depending on the model parameters and the mode k, the mass squared M^2 can be smaller than k^2/a^2 during the matter era [45].

In the regime $M^2/F \gg k^2/a^2$ the matter perturbation equation (182) reduces to the standard one in Einstein gravity, which gives the evolution $\delta_m \propto t^{2/3}$. For the model (98) the matter-induced mode of the field perturbation evolves as $\delta\phi_{\rm ind} \propto \delta\rho_m/M^2 \propto t^{\frac{2(4-p)}{3(1-p)}}$. Meanwhile, the WKB solution to (186) is given by $\delta\phi_{\rm osc} \propto t^{\frac{p}{2(1-p)}} \cos\left(ct^{-\frac{1}{1-p}}\right)$, where c is a constant. Since the background field ϕ during the matter era evolves as $\phi \propto t^{\frac{2}{1-p}}$, we find

$$\delta\phi/\phi = (\delta\phi_{\rm ind} + \delta\phi_{\rm osc})/\phi \simeq c_1 t^{2/3} + c_2 t^{-\frac{4-p}{2(1-p)}} \cos\left(ct^{-\frac{1}{1-p}}\right). \tag{187}$$

As long as the oscillating mode is initially suppressed relative to the matter-induced mode, the matter-induced mode remains the dominant contribution.

In the regime $M^2/F \ll k^2/a^2$ the effective gravitational constant is given by (184), which shows that the effect of modified gravity becomes important. Solving (182) in this case, we obtain the solution for matter perturbations

$$\delta_m \propto t^{\frac{\sqrt{25+48Q^2}-1}{6}}. \tag{188}$$

Setting $Q = -1/\sqrt{6}$, this recovers the solution $\delta_m \propto t^{(\sqrt{33}-1)/6}$ in $f(R)$ gravity.

The potential (98) has a heavy mass M which is much larger than H in the deep matter-dominated epoch, but it gradually decreases to become of the order of H around the present epoch. Depending on the modes k, the system crosses the point $M^2/F = k^2/a^2$ at $t = t_k$ during the matter era. Since for the model (98) M evolves as $M \propto t^{-\frac{2-p}{1-p}}$ during the matter era, the time t_k has a scale dependence given by $t_k \propto k^{-\frac{3(1-p)}{4-p}}$. When $t < t_k$ the evolution of δ_m is given by $\delta_m \propto t^{2/3}$, but for $t > t_k$ its evolution changes to the form given by (188).

During the matter era the mass squared is approximately given by

$$M^2 \simeq \frac{1-p}{(2^p\, p\, C)^{1/(1-p)}} Q^2 \left(\frac{\rho_m}{U_0}\right)^{\frac{2-p}{1-p}} U_0. \tag{189}$$

Using the relation $\rho_m = 3F_0\Omega_m^{(0)}H_0^2(1+z)^3$, we find that the critical redshift z_k at time t_k can be estimated as

$$z_k \simeq \left[\left(\frac{k}{a_0 H_0} \frac{1}{Q} \right)^{2(1-p)} \frac{2^p p C}{(1-p)^{1-p}} \frac{1}{(3F_0 \Omega_m^{(0)})^{2-p}} \frac{U_0}{H_0^2} \right]^{\frac{1}{4-p}} - 1, \qquad (190)$$

where a_0 is the present scale factor. The critical redshift increases for larger $k/(a_0 H_0)$. The matter power spectrum, in the linear regime, has been observed for the scales $0.01h\,\mathrm{Mpc}^{-1} < k < 0.2h\ \mathrm{Mpc}^{-1}$, which corresponds to $30a_0H_0 < k < 600a_0H_0$. In Fig. 4 we plot the evolution of the growth rate $s = \dot{\delta}_m/(H\delta_m)$ for the mode $k = 600a_0H_0$ and the coupling $Q = 1.08$ with three different values of p. Note that the asymptotic values of s in the regions $t \ll t_k$ and $t \gg t_k$ are given by $s = 1$ and $s = (\sqrt{25+48Q^2} - 1)/4$, respectively. We find that, for the scales $30a_0H_0 < k < 600a_0H_0$, the critical redshift exists in the region $z_k > 1$ and that z_k increases for smaller p. When $p = 0.7$ we have $z_k = 3.9$ from (190), which is consistent with the numerical result shown in Fig. 4. The growth rate s reaches a maximum value $s_{\max}$ and then begins to decrease around the end of the matter era.

The observational constraint on s reported by McDonald et al. [97] is $s = 1.46 \pm 0.49$ around the redshift $z = 3$, whereas the more recent data reported by Viel and Haehnelt [98] in the redshift range $2 < z < 4$ show that even the value $s = 2$ can be allowed in some of the observations. If we use the criterion $s < 2$ for the analytic estimation $s = (\sqrt{25+48Q^2}-1)/4$, we obtain the bound $Q < 1.08$. Figure 4 shows

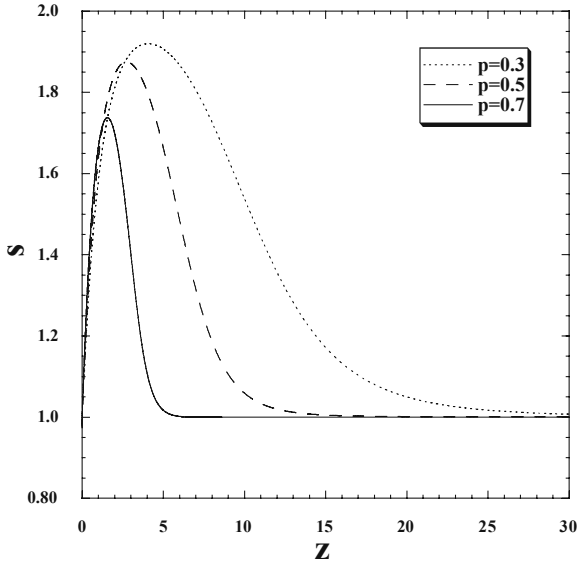

Fig. 4 Evolution of the growth rate s of matter perturbations in terms of the redshift z for $Q = 1.08$ and $k = 600a_0H_0$ with three different values of p. For smaller p the critical redshift z_k gets larger. The growth rate s reaches a maximum value and begins to decrease after the system enters the accelerated epoch. For smaller p the maximum value of s tends to approach the analytic value $(\sqrt{25+48Q^2} - 1)/4$

that $s_{\rm max}$ is smaller than the analytic value $s = 2$ (which corresponds to $Q = 1.08$). When $p = 0.7$, for example, we have that $s_{\rm max} = 1.74$. For the values of p that are very close to 1, $s_{\rm max}$ can be smaller than 1.5. However, these cases are hardly distinguishable from the ΛCDM model. In any case the current observational data on the growth rate s are not enough to place tight bounds on Q and p.

As in the case of $f(R)$ gravity, the matter power spectrum P_{δ_m} at time $t = t_\Lambda$ (at which $\ddot{a} = 0$) shows a difference compared to the ΛCDM model given by

$$\frac{P_{\delta_m}(t_\Lambda)}{P_{\delta_m}^{\Lambda{\rm CDM}}(t_\Lambda)} = \left(\frac{t_\Lambda}{t_k}\right)^{2\left(\frac{\sqrt{25+48Q^2}-1}{6}-\frac{2}{3}\right)} \propto k^{\frac{(1-p)(\sqrt{25+48Q^2}-5)}{4-p}}. \tag{191}$$

The CMB power spectrum is also modified by the non-standard evolution of the effective gravitational potential

$$\psi = \Phi - \Psi = \frac{3a^2H^2}{k^2}\Omega_m\delta_m, \tag{192}$$

which mainly affects the low multipoles because of the ISW effect. Since the smaller scale modes in CMB relevant to the galaxy power spectrum are hardly affected by this modification, there is a difference between the spectral indices of the matter power spectrum and of the CMB spectrum on the scales, $k > 0.01h\,{\rm Mpc}^{-1}$:

$$\Delta n(t_\Lambda) = \frac{(1-p)(\sqrt{25+48Q^2}-5)}{4-p}. \tag{193}$$

This reproduces the result (172) in $f(R)$ gravity by setting $Q = -1/\sqrt{6}$ and $p = 2n + 1$. If we use the criterion $\Delta n(t_\Lambda) < 0.05$, as in the case of the $f(R)$ gravity, we obtain the bounds $p > 0.957$ for $Q = 1$ and $p > 0.855$ for $Q = 0.5$. As long as p is close to 1, it is possible to satisfy both cosmological and local gravity constraints for $|Q| < 1$.

6.3 DGP Braneworld Model

In this section we discuss the evolution of linear matter perturbations in the DGP braneworld model. The perturbed metric in the 5D longitudinal gauge with four scalar metric perturbations Ψ, Φ, B, E is given by [56, 99]

$${\rm d}s^2 = -(1+2\Psi)n(t,y)^2{\rm d}t^2 + (1+2\Phi)A(t,y)^2\delta_{ij}{\rm d}x^i{\rm d}x^j + 2r_cB_{,i}{\rm d}x^i{\rm d}y + (1+2E){\rm d}y^2, \tag{194}$$

where the brane is located at $y = 0$ in the fifth dimension characterized by the coordinate y (we are considering a flat FLRW space-time on the brane). Note that B can be identified as a brane bending mode describing a perturbation of the brane

location and that r_c is the crossover scale defined in (131). The solution for the background metric describing the self-accelerating universe is [71]

$$n(t,y) = 1 + H(1 + \dot{H}/H^2)y, \qquad A(t,y) = a(t)(1 + Hy). \tag{195}$$

Recall that the Hubble parameter $H = \dot{a}/a$ satisfies (134) with $\varepsilon = +1$.

In the following we shall neglect the terms suppressed by the factor $aH/k \ll 1$ because we are considering sub-horizon perturbations. We also neglect the terms such as $(A'/A)\Phi'$, where a prime represents a derivative with respect to y. This comes from the fact that Φ' is of the order of $(k/a)\Phi$, as we will show later. The time-derivative terms can also be dropped under a quasi-static approximation. Then the perturbed 5D Einstein tensors $\delta\tilde{G}^A_B$ obey the following equations locally in the bulk [99]:

$$\delta\tilde{G}^0_0 = 3\Phi'' + \frac{2}{A^2}\nabla^2\Phi + \frac{\nabla^2}{A^2}(E - r_c B') - 2\frac{r_c}{A^2}\left(\frac{A'}{A}\right)\nabla^2 B = 0, \tag{196}$$

$$\delta\tilde{G}^i_j = -\frac{1}{A^2}(\nabla^i\nabla_j - \delta^i_j\nabla^2)(\Phi + \Psi + E - r_c B') + \delta^i_j(\Psi'' + 2\Phi'') + \frac{r_c}{A^2}(\nabla^i\nabla_j - \delta^i_j\nabla^2)\left(\frac{A'}{A} + \frac{n'}{n}\right)B = 0, \tag{197}$$

$$\delta\tilde{G}^5_i = -(\Psi' + 2\Phi')_{,i} = 0, \tag{198}$$

$$\delta\tilde{G}^5_5 = \frac{1}{A^2}\nabla^2(\Psi + 2\Phi) - \frac{r_c}{A^2}\left(2\frac{A'}{A} + \frac{n'}{n}\right)\nabla^2 B = 0. \tag{199}$$

Taking the divergence of the traceless part of (197), we get

$$\frac{\nabla^2}{A^2}(\Phi + \Psi + E - r_c B') - \frac{r_c}{A^2}\left(\frac{A'}{A} + \frac{n'}{n}\right)\nabla^2 B = 0. \tag{200}$$

For the consistency between (198) and (199), it is required that

$$B' = 0, \qquad \Psi' + 2\Phi' = 0. \tag{201}$$

From (199) and (200), we obtain

$$\frac{\nabla^2}{A^2}(E - r_c B') = -\frac{1}{2}\frac{\nabla^2}{A^2}\Psi + \frac{r_c}{2A^2}\frac{n'}{n}\nabla^2 B. \tag{202}$$

Substituting (199) and (202) into (196) together with the use of (201), we find

$$\Psi'' + \frac{\nabla^2}{A^2}\Psi - \frac{n'}{n}\frac{r_c}{A^2}\nabla^2 B = 0. \tag{203}$$

Under the sub-horizon approximation ($k/aH \gg 1$) the solution of (203), upon the Fourier transformation, is given by

$$\Psi - \frac{n'}{n} r_c B = \left[c_1 (1+Hy)^{-k/aH} + c_2 (1+Hy)^{k/aH} \right], \tag{204}$$

where c_1 and c_2 are integration constants. In order to avoid the divergence of the perturbation in the limit $y \to \infty$, we shall choose $c_2 = 0$.

The junction condition at the brane can be expressed in terms of an extrinsic curvature $K_{\mu\nu}$ and an energy momentum tensor on the brane:

$$K_{\mu\nu} - K g_{\mu\nu} = -\frac{\kappa_{(5)}^2}{2} T_{\mu\nu} + r_c G_{\mu\nu}, \tag{205}$$

where $K \equiv K^\mu_\mu$. Note that the extrinsic curvature is defined as $K_{\mu\nu} = h^\lambda_\mu \nabla_\lambda n_\nu$, where n_ν is the unit vector normal to the brane and $h_{\mu\nu} = g_{\mu\nu} - n_\mu n_\nu$ is the induced metric on the brane. The (0,0) and spatial components of the junction condition (205) give

$$\frac{2}{a^2}\nabla^2 \Phi = -\kappa_{(4)}^2 \delta\rho_m + \frac{1}{a^2}\nabla^2 B - \frac{3}{r_c}\Phi', \tag{206}$$

$$\Phi + \Psi = B, \tag{207}$$

$$\Psi' + 2\Phi' = 0, \tag{208}$$

where $\delta\rho_m$ is the matter perturbation on the brane. Equation (208) is consistent with the latter of (201).

From (204) it follows that $\Phi' \sim (k/a)\Phi$ in Fourier space. For the perturbations whose wavelengths are much smaller than the crossover scale r_c, i.e., $r_c\, k/a \gg 1$, we find that the term $(3/r_c)\Phi'$ in (206) is much smaller than $(k^2/a^2)\Phi$. In Fourier space (206) is approximately given by

$$\frac{2k^2}{a^2}\Phi = \kappa_{(4)}^2 \delta\rho_m + \frac{k^2}{a^2} B. \tag{209}$$

Using the projection of (199) as well as (207) and (209), we find that metric perturbations Ψ and Φ obey the following equations:

$$\frac{k^2}{a^2}\Psi = -\frac{\kappa_{(4)}^2}{2}\left(1 + \frac{1}{3\beta}\right)\delta\rho_m, \qquad \frac{k^2}{a^2}\Phi = \frac{\kappa_{(4)}^2}{2}\left(1 - \frac{1}{3\beta}\right)\delta\rho_m, \tag{210}$$

where

$$\beta(t) \equiv 1 - \frac{2r_c}{3}\left(2\frac{A'}{A} + \frac{n'}{n}\right) = 1 - 2Hr_c\left(1 + \frac{\dot{H}}{3H^2}\right). \tag{211}$$

The matter perturbation δ_m satisfies the same form of equation as given in (151) for the modes deep inside the horizon [55, 56]. Substituting the former of (210) into (151), we find that the matter perturbation obeys the following equation:

$$\ddot{\delta}_m + 2H\dot{\delta}_m - 4\pi G_{\rm eff}\rho_m\delta_m = 0, \tag{212}$$

where

$$G_{\rm eff} = \left(1 + \frac{1}{3\beta}\right) G. \tag{213}$$

Here G is 4D gravitational constant.

In the deep matter era one has $Hr_c \gg 1$ and hence $\beta \simeq -Hr_c$, so that β is largely negative ($|\beta| \gg 1$). In this regime the evolution of the matter perturbation is similar to that in General Relativity ($\delta_m \propto t^{2/3}$). The system finally approaches the de Sitter solution characterized by $H_{\rm dS} = 1/r_c$. We then have $\beta \simeq 1-2Hr_c \simeq -1$ around the de Sitter solution. Since $1 + 1/(3\beta) \simeq 2/3$, the growth rate in this regime is smaller relative to the case of General Relativity. The index γ of the growth rate $f = \Omega_m^{\gamma}$ is approximated by $\gamma \approx 0.68$ [100], which is different from the value $\gamma \simeq 0.55$ for the ΛCDM model. If the future imaging survey of galaxies can constrain γ within 20%, it may be possible to distinguish the ΛCDM model from DGP-modified gravity observationally [101].

Comparing (213) with the effective gravitational constant (184) in Brans–Dicke theory with a massless limit (or the absence of the field potential), we find that the Brans–Dicke parameter $\omega_{\rm BD}$ has the following relation with β:

$$\omega_{\rm BD} = \frac{3}{2}(\beta - 1). \tag{214}$$

Since $\beta < 0$ for the self-accelerating DGP solution, this implies that $\omega_{\rm BD} < -3/2$. This corresponds the theory with ghosts, because the kinetic energy of a scalar field degree of freedom is negative in the Einstein frame. Note that another normal branch of solutions in the DGP model does not suffer from this problem because the minus sign of (211) is replaced by the plus sign. In other words, the self-accelerating solution in the original DGP model can be realized at the expense of an appearance of the ghost state.

6.4 Observables in Weak Lensing

In the previous sections we have shown that modified gravity models generally lead to a change of the growth rate of matter perturbations compared to the ΛCDM model. Since there are two free functions that determine the first-order metrics Ψ and Φ, dark energy models can be classified according to how the gravitational potentials are linked to δ_m. In order to quantify this, we introduce two quantities $q(k,t)$ and $\zeta(k,t)$ defined by

$$\frac{k^2}{a^2}\Phi = 4\pi G q \delta_m \rho_m, \tag{215}$$

$$\frac{\Phi + \Psi}{\Phi} = \zeta, \tag{216}$$

where G is the 4D bare gravitational constant. Note that ζ characterizes the strength of the anisotropic stress. The ΛCDM model corresponds to $q = 1$ and $\zeta = 0$ (recall that the cosmological constant does not cluster). Modified gravity models give rise to different values of q,ζ relative to the ΛCDM model. Therefore, the functions q and ζ characterize a gravity theory for first-order scalar perturbations on small scales.

In the scalar-tensor model discussed in Sect. 6.2, the gravitational potentials are given by (181) on sub-horizon scales. In this case we have

$$q = \frac{1}{F}\frac{(k^2/a^2)(1-2Q^2)F + M^2}{(k^2/a^2)F + M^2}, \qquad \zeta = -\frac{4F(k^2/a^2)Q^2}{(k^2/a^2)(1-2Q^2)F + M^2}, \tag{217}$$

where we have used the unit $8\pi G = 1$. In the regime $M^2/F \gg k^2/a^2$ (and $F \simeq 1$) it follows that $q \simeq 1$ and $\zeta \simeq 0$. In the regime $M^2/F \ll k^2/a^2$ we have $q \simeq (1-2Q^2)/F$ and $\zeta \simeq -4Q^2/(1-2Q^2)$, so that the deviation from the ΛCDM model becomes important. Recall that the expression (217) covers the case of $f(R)$ gravity by setting $Q = -1/\sqrt{6}$. In [102] the quantities q and ζ have been evaluated for the more general Lagrangian density $f(R,\phi,X)$.

In the DGP model the gravitational potentials obey (210), which gives

$$q = 1 - \frac{1}{3\beta}, \qquad \zeta = \frac{2}{1-3\beta}. \tag{218}$$

In the deep matter era one has $|\beta| \gg 1$, so that $q \simeq 1$ and $\zeta \simeq 0$. The deviation from $(q,\zeta) = (1,0)$ appears when $|\beta|$ decreases to the order of unity, i.e., when the universe enters the epoch of late-time cosmic acceleration.

In order to confront dark energy models with the observations of weak lensing, it may be convenient to introduce the following quantity [93]:

$$\Sigma \equiv q(1 - \zeta/2). \tag{219}$$

From (215) and (216) we find that the weak lensing potential $\psi = \Phi - \Psi$ can be expressed as

$$\psi = 8\pi G \frac{a^2}{k^2}\rho_m \delta_m \Sigma. \tag{220}$$

We have $\Sigma = 1$ and $\Sigma = 1/F$ for the DGP model and for scalar-tensor models.

The effect of modified gravity theories manifests itself in weak lensing observations in at least two ways. One is the multiplication of the term Σ on the r.h.s. of (220). Another is the modification of the evolution of δ_m. The latter depends on two parameters q and ζ or, equivalently, on Σ and ζ. Thus two parameters (Σ,ζ) will be useful to detect signatures of modified gravity theories from future surveys of weak lensing.

7 Conclusions

We have discussed cosmological viability of modified gravity models as well as local gravity constraints on these models. In $f(R)$ gravity the coupling strength between dark energy and non-relativistic matter is of order 1 ($Q = -1/\sqrt{6}$) in the Einstein frame. Under the chameleon mechanism it is possible for $f(R)$ models to be consistent with local gravity constraints as long as the models are carefully designed. The models also need to possess a matter-dominated era followed by a late-time acceleration. It is required that viable $f(R)$ models satisfy the conditions $f_{,R} > 0$ and $f_{,RR} > 0$ for $R \geq R_1$, where R_1 is a Ricci scalar at a late-time de Sitter point. Moreover, for the consistency with local gravity tests, the variable $m = Rf_{,RR}/f_{,R}$ must approach 0 as R gets larger. In addition we require the stability condition (28) for the late-time de Sitter point. The representative models satisfying these requirements are given in (58), (59), and (60).

We have also considered a class of dark energy models based on scalar-tensor theories given by the action (77). In these theories, expressed in the Einstein frame, the scalar field ϕ couples to non-relativistic matter with a constant coupling Q. The action (77) is equivalent to the Brans–Dicke theory with a field potential U, where the Brans–Dicke parameter $\omega_{\rm BD}$ is related to the coupling Q via the relation $3 + 2\omega_{\rm BD} = 1/(2Q^2)$. This includes $f(R)$ gravity and quintessence models as special cases where the coupling is given by $Q = -1/\sqrt{6}$ (i.e., $\omega_{\rm BD} = 0$) and $Q = 0$ (i.e. $\omega_{\rm BD} \to \infty$), respectively. Even when $|Q|$ is of the order of 1, it is possible for Brans–Dicke models to be consistent with cosmological and local gravity constraints as long as the field potential is designed in a suitable way. One of the representative potentials satisfying these constraints is given in (98).

For viable $f(R)$ and Brans–Dicke models, we have shown that the quantity $F = \partial f/\partial R$ tends to increase from its present value F_0 as we go into the past, which results in the equation of state $w_{\rm DE}$ of dark energy becoming singular when $\Omega_m = F_0/F$. This can happen even around the redshift $z = 2$–3. This property is an important signature to distinguish these models from the ΛCDM cosmology.

We have studied the evolution of density perturbations for viable $f(R)$ and Brans–Dicke models. In the deep matter era a scalar-field degree of freedom has a large mass M to make these models compatible with local gravity constraints, but the mass gradually gets smaller as the universe enters the accelerated epoch. In the early cosmological epoch, there exists a "General Relativistic" phase during the matter era characterized by the condition $M^2 \gg k^2/a^2$ (k is a comoving wavenumber and

a is a scale factor). At this stage the matter perturbation evolve as $\delta_m \propto t^{2/3}$, as in the case of Einstein gravity. Around the end of the matter-dominated epoch, the deviation from Einstein gravity can be seen once M^2 becomes smaller than k^2/a^2. The evolution of perturbations during this "scalar-tensor" regime is given by $\delta_m \propto t^{\frac{\sqrt{25+48Q^2}-1}{6}}$. Under the criterion $s = \dot{\delta}_m/H\delta_m < 2$ for the growth rate of matter perturbations, we obtain the bound $Q < 1.08$ by using the analytic estimation $s = (\sqrt{25+48Q^2}-1)/4$. The difference Δn of the spectral indices of CMB and matter power spectra gives rise to another constraint on model parameters, e.g., $n \geq 2$ for the $f(R)$ models (59) and (60).

We also discussed the DGP braneworld model as a candidate for the late-time cosmic acceleration. While the universe exhibits a self-acceleration at late times, the joint constraints from data of SNLS, BAO, and the CMB shift parameter show that this model is under strong observational pressure. Moreover, this model contains a ghost mode with the effective Brans–Dicke parameter $\omega_{\rm BD}$ smaller than $-3/2$. Hence the original DGP model is effectively ruled out from observational constraints as well as from the ghost problem.

We have seen that modified gravity models generally give rise to a number of interesting observational signatures such as a divergence behavior of the dark energy equation of state as well as the peculiar evolution of matter density perturbations. We hope to find such deviations from the ΛCDM model in future high-precision observations.

References

1. V. Sahni and A. A. Starobinsky, Int. J. Mod. Phys. D **9**, 373 (2000). 100
2. S. M. Carroll, Living Rev. Rel. **4**, 1 (2001). 100
3. T. Padmanabhan, Phys. Rept. **380**, 235 (2003). 100
4. P. J. E. Peebles and B. Ratra, Rev. Mod. Phys. **75**, 559 (2003). 100
5. E. J. Copeland, M. Sami and S. Tsujikawa, Int. J. Mod. Phys. D **15**, 1753 (2006). 100
6. T. P. Sotiriou and V. Faraoni, arXiv:0805.1726 [gr-qc]. 100
7. Y. Fujii, Phys. Rev. D **26**, 2580 (1982). 100
8. L. H. Ford, Phys. Rev. D **35**, 2339 (1987). 100
9. C. Wetterich, Nucl. Phys B. **302**, 668 (1988). 100
10. B. Ratra and J. Peebles, Phys. Rev D **37**, 321 (1988). 100
11. R. R. Caldwell, R. Dave and P. J. Steinhardt, Phys. Rev. Lett. **80**, 1582 (1998). 100
12. T. Chiba, T. Okabe and M. Yamaguchi, Phys. Rev. D **62**, 023511 (2000). 100
13. C. Armendariz-Picon, V. F. Mukhanov and P. J. Steinhardt, Phys. Rev. Lett. **85**, 4438 (2000). 100
14. S. Capozziello, Int. J. Mod. Phys. **D 11**, 483, (2002). 100
15. S. Capozziello, V. F. Cardone, S. Carloni and A. Troisi, Int. J. Mod. Phys. **D**, 12, 1969 (2003). 100
16. S. M. Carroll, V. Duvvuri, M. Trodden and M. S. Turner, Phys. Rev. D **70**, 043528 (2004). 100
17. S. Nojiri and S. D. Odintsov, Phys. Rev. D **68**, 123512 (2003). 100
18. L. Amendola, Phys. Rev. D **60**, 043501 (1999). 100
19. J. P. Uzan, Phys. Rev. D **59**, 123510 (1999). 100
20. T. Chiba, Phys. Rev. D **60**, 083508 (1999). 100
21. N. Bartolo and M. Pietroni, Phys. Rev. D **61** 023518 (2000). 100
22. F. Perrotta, C. Baccigalupi and S. Matarrese, Phys. Rev. D **61**, 023507 (2000). 100
23. G. R. Dvali, G. Gabadadze and M. Porrati, Phys. Lett. B **485**, 208 (2000). 100, 101, 120

24. S. Nojiri, S. D. Odintsov and M. Sasaki, Phys. Rev. D **71**, 123509 (2005). 100, 126
25. T. Chiba, Phys. Lett. B **575**, 1 (2003). 100, 113
26. A. D. Dolgov and M. Kawasaki, Phys. Lett. B **573**, 1 (2003). 100
27. S. M. Carroll, I. Sawicki, A. Silvestri and M. Trodden, New J. Phys. **8**, 323 (2006). 100
28. R. Bean, D. Bernat, L. Pogosian, A. Silvestri and M. Trodden, Phys. Rev. D **75**, 064020 (2007). 100
29. Y. S. Song, W. Hu and I. Sawicki, Phys. Rev. D **75**, 044004 (2007). 100
30. I. Sawicki and W. Hu, Phys. Rev. D **75**, 127502 (2007). 100
31. L. Amendola, D. Polarski and S. Tsujikawa, Phys. Rev. Lett. **98**, 131302 (2007). 100
32. L. Amendola, D. Polarski and S. Tsujikawa, Int. J. Mod. Phys. D **16**, 1555 (2007). 100
33. G. J. Olmo, Phys. Rev. D **72**, 083505 (2005). 100
34. I. Navarro and K. Van Acoleyen, JCAP **0702**, 022 (2007). 100
35. T. Faulkner, M. Tegmark, E. F. Bunn and Y. Mao, Phys. Rev. D **76**, 063505 (2007). 100, 108
36. L. Amendola, R. Gannouji, D. Polarski and S. Tsujikawa, Phys. Rev. D **75**, 083504 (2007). 100, 102, 104, 1
37. B. Li and J. D. Barrow, Phys. Rev. D **75**, 084010 (2007). 100, 105
38. L. Pogosian and A. Silvestri, Phys. Rev. D **77**, 023503 (2008). 100
39. W. Hu and I. Sawicki, Phys. Rev. D **76**, 064004 (2007). 100, 106, 108, 110, 111
40. A. A. Starobinsky, JETP Lett. **86**, 157 (2007). 100, 110, 111, 129, 130
41. S. A. Appleby and R. A. Battye, Phys. Lett. B **654**, 7 (2007). 100, 110
42. S. Tsujikawa, Phys. Rev. D **77**, 023507 (2008). 100, 110, 129, 130, 132
43. C. Brans and R. H. Dicke, Phys. Rev. **124**, 925 (1961). 100, 112
44. L. Amendola, Phys. Rev. D **62**, 043511 (2000). 100
45. S. Tsujikawa, K. Uddin, S. Mizuno, R. Tavakol and J. Yokoyama, Phys. Rev. D **77**, 103009 (2008). 100, 113, 114, 116, 118, 133, 134, 135
46. S. Capozziello and S. Tsujikawa, Phys. Rev. D **77**, 107501 (2008). 100, 108, 111
47. P. Brax, C. van de Bruck, A. C. Davis and D. J. Shaw, Phys. Rev. D **78**, 104021 (2008). 100, 108
48. J. Khoury and A. Weltman, Phys. Rev. Lett. **93**, 171104 (2004). 100, 108, 109
49. J. Khoury and A. Weltman, Phys. Rev. D **69**, 044026 (2004). 100, 108, 109
50. I. Sawicki and S. M. Carroll, arXiv:astro-ph/0510364. 101, 125
51. M. Fairbairn and A. Goobar, Phys. Lett. B **642** , 432 (2006). 101, 125
52. R. Maartens and E. Majerotto, Phys. Rev. D **74**, 023004 (2006). 101, 125
53. U. Alam and V. Sahni, Phys. Rev. D **73**, 084024 (2006). 101, 125
54. Y. S. Song, I. Sawicki and W. Hu, Phys. Rev. D **75**, 064003 (2007). 101, 125
55. A. Lue, R. Scoccimarro and G. D. Starkman, Phys. Rev. D **69**, 124015 (2004). 101, 125, 140
56. K. Koyama and R. Maartens, JCAP **0601**, 016 (2006). 101, 125, 137, 140
57. D. Gorbunov, K. Koyama and S. Sibiryakov, Phys. Rev. D **73**, 044016 (2006). 101, 125
58. D. N. Vollick, Phys. Rev. D **68**, 063510 (2003). 101
59. X. Meng and P. Wang, Class. Quant. Grav. 20, 4949 (2003). 101
60. E. E. Flanagan, Phys. Rev. Lett. **92**, 071101 (2004). 101
61. T. P. Sotiriou, Class. Quant. Grav. **23**, 1253 (2006). 101
62. S. Tsujikawa, K. Uddin and R. Tavakol, Phys. Rev. D **77**, 043007 (2008). 101
63. A. A. Starobinsky, Phys. Lett. B **91**, 99 (1980). 102
64. L. Amendola and S. Tsujikawa, Phys. Lett. B **660**, 125 (2008). 104, 106
65. B. Boisseau, G. Esposito-Farese, D. Polarski and A. A. Starobinsky, Phys. Rev. Lett. **85**, 2236 (2000). 106, 133
66. S. Tsujikawa, Phys. Rev. D **77**, 023507 (2008). 106
67. K. i. Maeda, Phys. Rev. D **39**, 3159 (1989). 106
68. T. Tamaki and S. Tsujikawa, Phys. Rev. D **78**, 084028 (2008). 108, 109
69. C. M. Will, Living Rev. Rel. **9**, 3 (2005). 109
70. G. R. Dvali and G. Gabadadze, Phys. Rev. D **63** (2001), 065007. 120
71. C. Deffayet, Phys. Lett. B **502** (2001), 199. 120, 121, 122, 123, 138
72. C. Deffayet, G. R. Dvali and G. Gabadadze, Phys. Rev. D **65** (2002), 044023. 120, 123
73. P. Binetruy, C. Deffayet and D. Langlois, Nucl. Phys. B **565**, 269 (2000). 121, 122

74. P. Binetruy, C. Deffayet, U. Ellwanger and D. Langlois, Phys. Lett. B **477**, 285 (2000). 121, 122
75. T. Shiromizu, K. i. Maeda and M. Sasaki, Phys. Rev. D **62**, 024012 (2000). 123
76. S. M. Carroll, A. De Felice, V. Duvvuri, D. A. Easson, M. Trodden and M. S. Turner, Phys. Rev. D **71**, 063513 (2005). 125
77. A. Nunez and S. Solganik, Phys. Lett. B **608**, 189 (2005). 126
78. G. Calcagni, S. Tsujikawa and M. Sami, Class. Quant. Grav. **22**, 3977 (2005). 126
79. A. De Felice, M. Hindmarsh and M. Trodden, JCAP **0608**, 005 (2006). 126
80. G. Calcagni, B. de Carlos and A. De Felice, Nucl. Phys. B **752**, 404 (2006). 126
81. T. Koivisto and D. F. Mota, Phys. Lett. B **644**, 104 (2007). 126
82. S. Tsujikawa and M. Sami, JCAP **0701**, 006 (2007). 126
83. T. Koivisto and D. F. Mota, Phys. Rev. D **75**, 023518 (2007). 126
84. B. M. Leith and I. P. Neupane, JCAP **0705**, 019 (2007). 126
85. Z. K. Guo, N. Ohta and S. Tsujikawa, Phys. Rev. D **75**, 023520 (2007). 126
86. L. Amendola, C. Charmousis and S. C. Davis, JCAP **0612**, 020 (2006). 126
87. S. Nojiri and S. D. Odintsov, Phys. Lett. B **631**, 1 (2005). 126
88. B. Li, J. D. Barrow and D. F. Mota, Phys. Rev. D **76**, 044027 (2007). 126
89. A. De Felice and S. Tsujikawa, Phys. Lett. B **675**, 1 (2009). 126
90. S. Y. Zhou, E. J. Copeland and P. M. Saffin, arXiv:0903.4610 [gr-qc]. 126
91. H. Kodama and M. Sasaki, Prog. Theor. Phys. Suppl. **78**, 1 (1984). 127
92. J. C. Hwang and H. Noh, Phys. Rev. D **71**, 063536 (2005). 127, 128, 133
93. L. Amendola, M. Kunz and D. Sapone, JCAP **0804**, 013 (2008). 132, 141
94. Y. S. Song, H. Peiris and W. Hu, Phys. Rev. D **76**, 063517 (2007). 132
95. S. Tsujikawa and T. Tatekawa, Phys. Lett. B **665**, 325 (2008). 132
96. F. Schmidt, Phys. Rev. D **78**, 043002 (2008). 132
97. P. McDonald et al., astro-ph/0407377. 136
98. M. Viel and M. G. Haehnelt, Mon. Not. Roy. Astron. Soc. **365**, 231 (2006). 136
99. K. Koyama and F. P. Silva, Phys. Rev. D **75**, 084040 (2007). 137, 138
100. E. V. Linder, Phys. Rev. D **72**, 043529 (2005). 140
101. K. Yamamoto, D. Parkinson, T. Hamana, R. C. Nichol and Y. Suto, Phys. Rev. D **76**, 023504 (2007). 140
102. S. Tsujikawa, Phys. Rev. D **76**, 023514 (2007). 141

Statistical Methods in Cosmology

L. Verde

Summary The advent of large data set in cosmology has meant that in the past 10 or 20 years our knowledge and understanding of the Universe has changed not only quantitatively but also, and most importantly, qualitatively. Cosmologists are interested in studying the origin and evolution of the physical Universe. They rely on data where a host of useful information is enclosed, but is encoded in a non-trivial way. The challenges in extracting this information must be overcome to make the most of the large experimental effort. Even after having analyzed a decade or more of data and having converged to a standard cosmological model (the so-called and highly successful ΛCDM model) we should keep in mind that this model is described by 10 or more physical parameters and if we want to study deviations from the standard model the number of parameters is even larger. Dealing with such a high-dimensional parameter space and finding parameters constraints is a challenge on itself. In addition, as gathering data is such an expensive and difficult process, cosmologists want to be able to compare and combine different data sets both for testing for possible disagreements (which could indicate new physics) and for improving parameter determinations. Finally, because experiments are always so expansive, cosmologists in many cases want to find out a priori, before actually doing the experiment, how much one would be able to learn from it. For all these reasons, more and more sophisticated statistical techniques are being employed in cosmology, and it has become crucial to know some statistical background to understand recent literature in the field. Here, I will introduce some statistical tools that any cosmologist should know about in order to be able to understand recently published results from the analysis of cosmological data sets. I will not present a complete and rigorous introduction to statistics as there are several good books which are reported in the references. The reader should refer to those. I will take a practical approach and I will touch upon useful tools such as statistical inference, Bayesians vs Frequentists approach, chisquare and goodness of fit, confidence regions, likelihood, Fisher matrix approach, Monte Carlo methods, and a brief introduction to model testing. Throughout, I will use practical examples often taken from recent

L. Verde (✉)
ICREA & ICE (IEEC-CSIC) and ICC UB, Bellaterra, Spain
e-mail: verde@ieec.uab.es

Verde, L.: *Statistical Methods in Cosmology*. Lect. Notes Phys. **800**, 147–177 (2010)
DOI 10.1007/978-3-642-10598-2_4

literature to illustrate the use of such tools. Of course this will not be an exhaustive guide: it should be interpreted as a "starting kit," and the reader is warmly encouraged to read the references to find out more.

1 Introduction

As cosmology has made the transition from a data-starved science to a data-driven science, the use of increasingly sophisticated statistical tools has increased. As explained in detail below, cosmology is intrinsically related to statistics, as theories of the origin and evolution of the Universe do not predict, for example, that a particular galaxy will form at a specific point in space and time or that a specific patch of the cosmic microwave background will have a given temperature; any theory will predict average statistical properties of our Universe, and we can only observe a particular realization of that.

It is often said that cosmology has entered the precision era: "precision" requires a good knowledge of the error bars and thus confidence intervals of a measurement. This is an inherently statistical statement. We should try, however, to go even further, and also achieve "accuracy" (although cosmology does not have a particularly stellar track record in this regard). This requires quantifying systematic errors (beyond the statistical ones) and it also requires statistical tools. For all these reasons, knowledge of basic statistical tools has become indispensable to understand the recent cosmological literature.

Examples of applications where probability and statistics are crucial in Cosmology are (i) Is the Universe homogenous and isotropic on large scales? (ii) Are the initial conditions consistent with being Gaussian? (iii) Is there a detection of non-zero tensor modes? (iv) What is the value of the density parameter of the Universe Ω_m given the WMAP data for a ΛCDM model? (v) What are the allowed values at a given confidence level for the primordial power spectrum spectral slope n? (vi) What is the best fit value of the dark energy equation of state parameter w? (vii) Is a model with equation of state parameter different from -1 a better fit to the data than a model with non-zero curvature? (viii) What will be the constraint on the parameter w for a survey with given characteristic?

The first three questions address the hypothesis-testing issue. You have an hypothesis and you want to check whether the data are consistent with it. Sometimes, especially for addressing issues of "detection" you can test the null hypothesis: assume the quantity is zero and test whether the data are consistent with it.

The next three questions are "parameter estimation" problems: we have a model, in this example, the ΛCDM model, which is characterized by some free parameters which we would like to measure.

The next question, (vii), belongs to "model testing"; we have two models and ask which one is a better fit to the data. Model testing comes in several different flavors: the two models to be considered may have different number of parameters or equal number of parameters, may have some parameters in common or not, etc.

Finally question (viii) is on "forecasting," which is particularly useful for or quickly forecasting the performance of future experiments and for experimental design.

Here we will mostly concentrate on the issue of parameter estimation but also touch upon the other applications.

2 Bayesians vs Frequentists

The world is divided into Frequentists and Bayesians. For Frequentists probabilities $\mathcal{P}$ are frequencies of occurrence:

$$\mathcal{P} = \frac{n}{N}, \tag{1}$$

where n denotes the number of successes and N the total number of trials. Frequentists define probability as the limit for the number of independent trials going to infinity. Bayesians interpret probabilities as *degree of belief in a hypothesis*.

Let us say that x is our random variable (event). Depending on the application, x can be the number of photons hitting a detector, the matter density in a volume, the Cosmic Microwave Background temperature in a direction in the sky, etc. The probability that x takes a specific value is $\mathcal{P}(x)$ where $\mathcal{P}$ is called probability distribution. Note that probabilities (the possible values of x) can be discrete or continuous. $\mathcal{P}(x)$ is a *probability density*: $\mathcal{P}(x)dx$ is the probability that the random variable x takes a value between x and $x + dx$. Frequentists only consider probability distributions of events while Bayesians consider hypothesis as events.

For both, the rules of probability apply.

1. $\mathcal{P}(x) \geq 0$
2. $\int_{-\infty}^{\infty} dx\mathcal{P}(x) = 1$. In the discrete case $\int \longrightarrow \sum$.
3. For mutually exclusive events $\mathcal{P}(x_1 U x_2) \equiv \mathcal{P}(x_1 .OR. x_2) = \mathcal{P}(x_1) + \mathcal{P}(x_2)$
4. In general $\mathcal{P}(x_1,x_2) = \mathcal{P}(x_1)\mathcal{P}(x_1|x_2)$. In words, the probability of x_1 AND x_2 to happen is the probability of x_1 times the *conditional probability* of x_2 given that x_1 has already happened.

The last item deserves some discussion. For example, only for independent events where $\mathcal{P}(x_2|x_1) = \mathcal{P}(x_2)$ one can write $\mathcal{P}(x_1,x_2) = \mathcal{P}(x_1)\mathcal{P}(x_2)$. Of course in general one can always rewrite $\mathcal{P}(x_1,x_2) = \mathcal{P}(x_1)\mathcal{P}(x_1|x_2)$ by switching $x_1$and x_2. If then one makes the apparently tautological identification that $\mathcal{P}(x_1,x_2) = \mathcal{P}(x_2,x_1)$ and substitute $x_1 \longrightarrow D$ standing for *data* and $x_2 \longrightarrow H$ standing for *hypothesis*, one gets Bayes theorem :

$$\mathcal{P}(H|D) = \frac{\mathcal{P}(H)\mathcal{P}(D|H)}{\mathcal{P}(D)}, \tag{2}$$

where $\mathcal{P}(H|D)$ is called the *posterior*, $\mathcal{P}(D|H)$ is the *likelihood* (the probability of the data given the hypothesis) and $\mathcal{P}(H)$ is called the *prior*. Note that here explicitly we have probability and probability distribution of a hypothesis.

3 Bayesian Approach and Statistical Inference

Despite its simplicity, Bayes theorem is at the base of statistical inference. For the Bayesian point of view let us use D to indicate our data (or data set). The hypothesis H can be a model, say for example the ΛCDM model, which is characterized by a set of parameters $\boldsymbol{\theta}$. In the Bayesian framework what we want to know is "What is the probability distribution for the model parameters given the data?" i.e. $\mathcal{P}(\boldsymbol{\theta}|D)$. From this information we can extract the most likely value for the parameters and their confidence limits.[1] However, what we can compute accurately, in most instances, is the likelihood, which is related to the posterior by the prior. (At this point one assumes that one has collected the data and so $\mathcal{P}(D) = 1$). The prior, however, can be somewhat arbitrary. This is a crucial point to which we will return below. For now let us consider an example: the constraint from WMAP data on the integrated optical depth to the last scattering surface τ. One could do the analysis using the variable τ itself, however, one could also note that the temperature data (the angular power spectrum of the temperature fluctuations) on large scale depend approximately linearly on the variable $Z = \exp(-2\tau)$. A third person would note that the polarization data (in particular the EE angular power spectrum) depend roughly linearly on τ^2. So person one could use a uniform prior in τ, person two a uniform prior in $\exp(-2\tau)$, and person three in τ^2. What is the relation between $\mathcal{P}(\tau)$, $\mathcal{P}(Z)$, and $\mathcal{P}(\tau^2)$?

3.1 Transformation of Variables

We wish to transform the probability distribution of $\mathcal{P}(x)$ to the probability distribution of $\mathcal{G}(y)$ with y as a function of x. Recall that probability is a conserved quantity (we cannot create or destroy probabilities . . .) so

$$\mathcal{P}(x)dx = \mathcal{G}(y)dy, \tag{3}$$

thus

$$\mathcal{P}(x) = \mathcal{G}(y(x)) \left| \frac{dy}{dx} \right|. \tag{4}$$

Following the example above if x is τ and y is $\exp(\tau)$ then $\mathcal{P}$ is related to $\mathcal{G}$ by a factor 2τ and if y is τ^2 by a factor 2. In other words using different priors leads to different posteriors. This is the main limitation of the Bayesian approach.

[1] At this point many Frequentists stop reading this document . . .

3.2 Marginalization

So far we have considered probability distributions of a random variable x, but one could analogously define *multi-variate distributions*, the joint probability distribution of two or more variables, e.g., $\mathcal{P}(x,y)$. A typical example is the description of the initial distribution of the density perturbations in the Universe. Motivated by inflation and by the central limit theorem, the initial distribution of density perturbation is usually described by a multi-variate Gaussian: at every point in space given by its spatial coordinates (x, y, z), $\mathcal{P}$ is taken to be a random Gaussian distribution. Another example is when one simultaneously constrains the parameters of a model, say, for example, $\boldsymbol{\theta} = \{\Omega_m, H_0\}$ (here H_0 denotes the Hubble constant). If you have $\mathcal{P}(\Omega_m,H_0)$ and want to know the probability distribution of Ω_m regardless of the values of H_0 then

$$\mathcal{P}(\Omega_m) = \int dH_0 \mathcal{P}(\Omega_m,H_0). \tag{5}$$

3.3 Back to Statistical Inference and Cosmology

Let us go back to the issue of statistical inference and follow the example from [1]. If you have an urn with N red balls and M blue balls and you draw one ball at the time then probability theory can tell you what are your chances of picking a red ball given that you have already drawn n red and m blue: $\mathcal{P}(D|H)$. However, this is not what you want to do, you want to make a few draws from the urn and use probability theory to tell you what is the red vs blue distribution inside the urn is, $\mathcal{P}(H|D)$. In the Frequentist approach all you can compute is $\mathcal{P}(D|H)$.

In the case of cosmology it gets even more complicated.

We consider that the Universe we live in is a random realization of all the possible Universes that could have been a realization of the true underlying model (which is known only to Mother Nature). All the possible realizations of this true underlying Universe make up the *ensemble*. In statistical inference one may sometime want to try to estimate how different our particular realization of the Universe could be from the true underlying one. Going back to the example of the urn with red and blue balls, it would as if we were to be drawing from one particular urn, but the urn is part of a large batch. On average, the batch distribution has 50% red and 50% blue, but each urn has only an odd number of balls and so any particular urn cannot reflect exactly the 50–50 spilt.

A crucial assumption of standard cosmology is that the part of the Universe that we can observe is a fair sample of the whole. But the peculiarity in cosmology is that we have just one Universe, which is just one realization from the ensemble (quite fictitious one: it is the ensemble of all possible Universes). The fair sample hypothesis states that samples from well-separated parts of the Universe are independent realizations of the same physical process, and that, in the observable part of the Universe, there are enough independent samples to be representative of the statistical ensemble.

In addition, experiments in cosmology are not like lab experiments: in many cases observations cannot be easily repeated (think about the observation of a particular supernova explosion or of a Gamma ray burst) and we cannot try to perturb the Universe to see how it reacts... After these considerations, it may be clearer why cosmologists tend to use the Bayesian approach.

4 Chisquare and Goodness of Fit

Say that you have a set of observations and have a model, described by a set of parameters $\boldsymbol{\theta}$, and want to fit the model to the data. The model may be physically motivated or a convenient function. One then should define a merit function, quantifying the agreement between the model and the data, by maximizing the agreement one obtains the best fit parameters. Any useful fitting procedure should provide: (1) best fit parameters (2) estimation of error on the parameters (3) possibly a measure of the goodness of fit. One should bear in mind that if the model is a poor fit to the data then the recovered best fit parameters are meaningless.

Following numerical recipes ([2], Chap. 15) we introduce the concept of model fitting (parameter fitting) using least squares. Let us assume we have a set of data points D_i, for example, these could be the band power galaxy power spectrum at a set of k values, and a model for these data $y(x, \boldsymbol{\theta})$ which depends on set of parameters $\boldsymbol{\theta}$ (e.g., the ΛCDM power spectrum, which depends on n_s-primordial power spectrum spectral slope, σ_8-present-day amplitude of rms mass fluctuations on scale of 8 Mpc/h–, $\Omega_m h$, etc.). Or it could be, for example, the supernovae type 1a distance modulus as a function of redshift; see, e.g., Fig. 1 [3, 4].

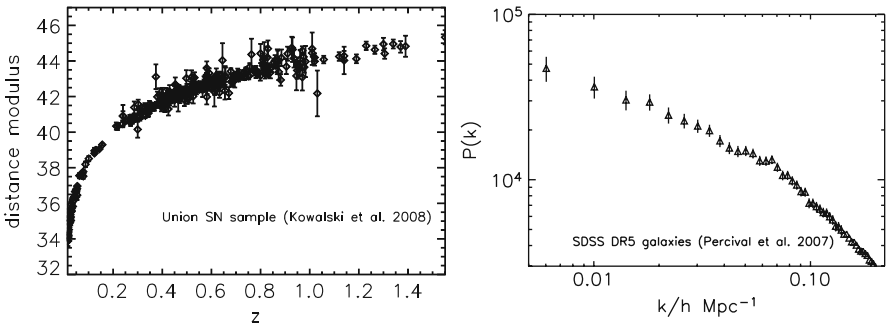

Fig. 1 *Left*: distance modulus vs redshift for supernovae type 1A from the UNion sample [3]. *Right*: bandpower $P(k)$ for DR5 SDSS galaxies, from [4]. In both cases one may fit a theory (and the theory parameters) to the data with the chisquare method. Note that in both cases errors are correlated. In the *right* panel the errors are also strictly speaking non-Gaussianly distributed

The least squares, in its simplest incarnation is

$$\chi^2 = \sum_i w_i [D_i - y(x_i|\boldsymbol{\theta})]^2, \tag{6}$$

where w_i are suitably defined weights. It is possible to show that the minimum variance weight is $w_i = 1/\sigma_i^2$ where σ_i denotes the error on data point i. In this case the least squares is called chisquare. If the data are correlated the chisquare becomes

$$\chi^2 = \sum_{ij} (D_i - y(x_i|\theta))\, Q_{ij} (D_j - y(x_j|\theta)), \tag{7}$$

where Q denotes the inverse of the so-called covariance matrix describing the covariance between the data. The best fit parameters are those that minimize the χ^2. See an example in Fig. 2.

For a wide range of cases the probability distribution for different values of χ^2 around the minimum of (7) is the χ^2 distribution for $\nu = n - m$ degrees of freedom where n is the number of independent data points and m the number of parameters. The probability that the observed χ^2 exceeds by chance a value $\widehat{\chi}$ for the correct model is $Q(\nu,\widehat{\chi}) = 1 - \Gamma(\nu/2,\widehat{\chi}/2)$ where Γ denotes the incomplete Gamma function. See the Numerical Recipes bible [2]. Conversely, the probability that the observed χ^2, even for the correct model, is less than $\widehat{\chi}$ is $1 - Q$. While this statement is strictly true if measurement errors are Gaussian and the model is a linear function of the parameters, in practice it applies to a much wider range of cases.

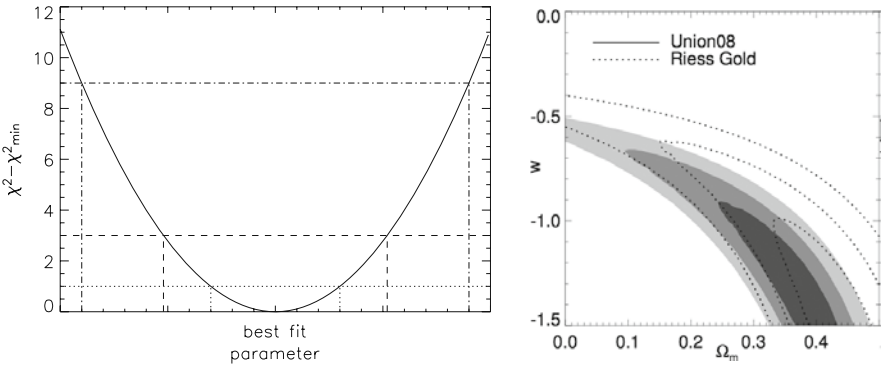

Fig. 2 *Left*: example of a one-dimensional chisquare for a Gaussian distribution as a function of a parameter and corresponding 68.3, 95.4, and 99.5% confidence levels. *Right*: a two-dimensional example for the union supernovae data. Figure from Kowalski et al. [3] reproduced with permission from the AAS. Note that in a practical application even if the data have Gaussian errors the errors on the parameter may not be well described by multi-variate Gaussians (thus the confidence regions are not ellipses)

The quantity Q evaluated that the minimum chisquare (i.e., at the best fit values for the parameters) gives a measure of the goodness of fit. If Q gives a very small probability then there are three possible explanations:

(1) the model is wrong and can be rejected. (Strictly speaking, the data are unlikely to have happened if the Universe was really described by the model considered)
(2) the errors are underestimated
(3) the measurement errors are non-Gaussianly distributed.

Note that in the example of the power spectrum we know a priori that the errors are non-Gaussianly distributed. In fact, even if the initial conditions were Gaussian and if the underlying matter perturbations were still evolving in the linear regime (i.e., $\delta\rho/\rho \ll 1$) and galaxies were nearly unbiased tracers of the dark matter, then the density fluctuation itself would obey Gaussian statistics and so would its Fourier transform, but *not* its power spectrum, which is a square quantity. In reality we know that by $z = 0$ perturbations grow non-linearly and that galaxies may not be nearly unbiased tracers of the underlying density field. Nevertheless, the central limit theorem comes to our rescue, if in each band power there is a sufficiently large number of modes.

If Q is too large (too good to be true) it is also cause for concern:

(1) errors have been overestimated
(2) data are correlated or non-independent
(3) the distribution is non-Gaussian

Beware: this last case is very rare.

A useful "chi-by-eye" rule is the minimum χ^2 should be roughly equal to ν (number of data–number of parameters). This is increasingly true for large ν. From this, it is easy to understand the use of the so-called reduced chisquare that is the $\chi^2_{\rm min}/m$: if $m \gg n$ (i.e., number of data much larger than the number of parameters to fit, which should be true in the majority of the cases) then $m \sim \nu$ and the rule of thumb is that reduced chisquare should be unity.

Note that the chisquare method, and the Q statistic, gives the probability for the data, given a model $\mathcal{P}(D|\theta)$ and not $\mathcal{P}(\theta|D)$. One can make this identification via the prior.

5 Confidence Regions

Once the best fit parameters are obtained, how can one represent the confidence limit or confidence region around the best fit parameters? A reasonable choice is to find a region in the m-dimensional parameter space (remember that m is the number of parameters) that contains a given percentage of the probability distribution. In most cases one wants a compact region around the best fit values. A natural choice is then given by regions of constant χ^2 boundaries. Note that there may be cases (when the χ^2 has more than one minimum) in which one may need to report a non-connected

confidence region. For multi-variate Gaussian distributions, however, these are ellipsoidal regions. Note that the fact that the data have Gaussian errors does not imply that the parameters will have a Gaussian probability distribution . . .

Thus, if the values of the parameters are perturbed from the best fit, the χ^2 will increase. One can use the properties of the χ^2 distribution to define confidence intervals in relation to χ^2variations or $\Delta\chi^2$. Table 1 reports the $\Delta\chi^2$ for 68.3, 95.4, and 99.5% confidence levels as function of number of parameters for the joint confidence level. In the case of Gaussian distributions these correspond to the conventional 1, 2, and 3σ. See an example of this in Fig. 2

Beyond these values here is the general prescription to compute constant χ^2 boundaries confidence levels. After having found the best fit parameters by minimizing the χ^2 and if Q for the best fit parameters is acceptable then

- Let m be the number of parameters, n the number of data, and p be the confidence limit
- Solve the following equation for $\Delta\chi^2$:

$$Q(n-m, \min(\chi^2) + \Delta\chi^2) = p \tag{8}$$

- Find the parameter region where $\chi^2 \leq \min(\chi^2) + \Delta\chi^2$. This defines the confidence region.

If the actual error distribution is non-Gaussian but it is known then it is still possible to use the χ^2 approach, but instead of using the chisquare distribution and Table 1, the distribution needs to be calibrated on multiple simulated realization of the data as illustrated below in Sect. 13.

Table 1 $\Delta\chi^2$ for the conventionals 1, 2, and $3-\sigma$ as a function of the number of parameters for the joint confidence levels

p (%)	1	2	3
68.3	1.00	2.30	3.53
95.4	2.71	4.61	6.25
99.73	9.00	11.8	14.2

6 Likelihood

So far we have dealt with the frequentist quantity $\mathcal{P}(D|H)$. If we set $\mathcal{P}(D) = 1$ and ignore the prior then we can identify the likelihood with $\mathcal{P}(H|D)$ and thus by maximizing the likelihood we can find the most likely model (or model's parameters) given the data. However, having ignored $\mathcal{P}(D)$ and the prior this approach cannot give in general a goodness of fit and thus cannot give an absolute probability for a

given model. However, it can give relative probabilities. If the data are Gaussianly distributed the likelihood is given by a multi-variate Gaussian:

$$\mathcal{L} = \frac{1}{(2\pi)^{n/2}|detC|^{1/2}} \exp\left[-\frac{1}{2}\sum_{ij}(D-y)_i C_{ij}^{-1}(D-y)_j\right], \tag{9}$$

where $C_{ij} = \langle (D_i - y_i)(D_j - y_j)\rangle$ is the covariance matrix.

It should be clear from this that the relation between χ^2 and likelihood is that, for Gaussian distributions, $\mathcal{L} \propto \exp[-1/2\chi^2]$ and minimizing the χ^2 is equivalent to minimizing the likelihood. In this case likelihood analysis and χ^2 coincide and by the end of this section, it will thus be no surprise that the Gamma function appearing in the χ^2 distribution is closely related to the Gaussian integrals.

The subtle step is that now, in Bayesian statistics, confidence regions are regions R in *model space* such that $\int_R \mathcal{P}(\theta|D)d\theta = p$ where p is the confidence level we request (e.g., 68.3, 95.4%). Note that by integrating the posterior over the model parameters, the confidence region depends on the prior information, as seen in Sect. 3.1 different priors give different posteriors and thus different regions R.

It is still possible to report results independently of the prior by using the *likelihood ratio*. The likelihood at a particular point in parameter space is compared with that at the best fit value, $\mathcal{L}_{\max}$, where likelihood is maximized. Thus a model is acceptable if the likelihood ratio,

$$\Lambda = -2\ln\left[\frac{\mathcal{L}(\theta)}{\mathcal{L}_{\max}}\right], \tag{10}$$

is above a given threshold. The connection to the χ^2 for Gaussian distribution should be clear. In general, the threshold can be calibrated by calculating the entire distribution of the likelihood ratio in the case that a particular model is the true model. Frequently this is chosen to be the best fit model.

There is a subtlety to point out here. In cosmology the data may be Gaussainly distributed and still the χ^2 and likelihood ratio analysis may give different results. This happens because in identifying likelihood and chisquare we have neglected the term $[(2\pi)^{n/2}|detC|^{1/2}]^{-1}$. If the covariance does not depend on the model or model parameters, this is just a normalization factor which drops out in the likelihood ratio. However, in cosmology often the covariance depends on the model; this happens, for example, when errors are dominated by cosmic variance, like in the case of the CMB temperature fluctuations on the largest scales, or on the galaxies power spectrum on the largest scales. In this case the cosmology dependence of the covariance cannot be neglected, but one can always define a pseudo-chisquare as $-2\ln\mathcal{L}$ and work with this quantity.

Let us stress again that the likelihood is linked to the posterior through the prior; the identification of the likelihood with the posterior is prior dependent (as we will see in an example below). In the absence of any data it is common to assume a flat (uniform) prior, i.e., all values of the parameter in question are equally likely, but

other choices are possible and sometimes more motivated. For example, if a parameter is positive definite, it may be interesting to use a logarithmic prior (uniform in the log).

Priors may be assigned theoretically or from prior information gathered from previous experiments. If the priors are set by theoretical considerations, it is always a good practice to check how much the results depend on the choice of the prior. If the dependence is significant, it means that the data do not have much statistical power to constrain that (those) parameter(s). Information theory helps us quantify the amount of "information gain"; the information in the posterior relative to the prior is

$$\mathcal{I} = \int \mathcal{P}(\theta|D) \log \left[\frac{\mathcal{P}(\theta|D)}{\mathcal{P}(\theta)} \right] d\theta. \tag{11}$$

6.1 Marginalization: Examples

Some of the model parameters may be uninteresting. For example, in many analyses one wants to include nuisance parameters (calibration factors, biases, etc.) but then report the confidence level on the real cosmological parameters regardless of the value of the nuisance ones. In other cases the model may have say, 10 or more real cosmological parameters but we may be interested in the allowed range of only one or two of them, regardless of the values of all the others. Typical examples are, e.g., constraints on the curvature parameter Ω_k (which we may want to know regardless of the values of, e.g., Ω_m or Ω_Λ) or, say, the allowed range for the neutrino mass regardless of the power spectrum spectral index or the value of the Hubble constant. As explained in Sect. 3.2 one can marginalize over the uninteresting parameters.

It should be kept in mind that marginalization is a Bayesian concept: the results may depend on the prior chosen.

In some cases, the marginalization can be carried out analytically. An example is reported below, this applies to the case of, e.g., calibration uncertainty, point sources amplitude, overall scale independent galaxy bias, magnitude intrinsic brightness, or beam errors for CMB studies. In this case it is useful to know the following results for Gaussian likelihoods:

$$\begin{aligned} \mathcal{P}(\theta_1..\theta_{m-1}|D) = \int & \frac{dA}{(2\pi)^{\frac{m}{2}} ||C||^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(C_i-(\hat{C}_i+AP_i))\Sigma^{-1}_{ij}(C_j-(\hat{C}_j+AP_j))\right]} \\ & \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(A-\hat{A})^2}{\sigma^2} \right], \end{aligned} \tag{12}$$

where repeated indices are summed over and $||C||$ denotes the determinant. Here, A is the amplitude of, say, a point source contribution P to the C_ℓ angular power spectrum, A is the mth parameter which we want to marginalize over with a Gaussian

prior with variance σ^2 around $\hat{A}$. The trick is to recognize that this integral can be written as

$$\mathcal{P}(\theta_1..\theta_{m-1}|D) = C_0 \exp\left[-\frac{1}{2}C_1 - 2C_2A + C_3A^2\right] dA, \tag{13}$$

(where $C_{0...3}$ denote constants and it is left as an exercise to write them down explicitly) and that this kind of integral is evaluated by using the substitution $A \longrightarrow A - C_2/C_3$ giving something $\propto \exp[-1/2(C_1 - C_2^2/C_3)]$.

In cases where the likelihood surface (describing the value of the likelihood as a function of the parameters) is not a multi-variate Gaussian, the location of the maximum likelihood before marginalization may not coincide with the location after marginalization. An example is shown in Fig. 3. The figure shows the probability distribution for Ω_k from WMAP5 data for a model where curvature is free and the equation of state parameter for dark energy w is constant in time but not fixed at -1. The red line shows the N-dimensional maximum posterior value and the black line is the marginalized posterior over all other cosmological parameters.

It should also be added that, even in the case where we have a single-peaked posterior probability distribution there are two common estimators of the "best" parameters: the peak value (i.e., the most probable value) or the mean, $\hat{\theta} = \int d\theta\theta\mathcal{P}(\theta|D)$. If the posterior is non-Gaussian these two estimates need not coincide. In the same spirit, slightly different definitions of confidence intervals need not coincide for non-Gaussian likelihoods, as illustrated in the right panel of Fig. 3: for example, one can define the confidence interval $[\theta_{low}, \theta_{high}]$, such that equal fractions of the

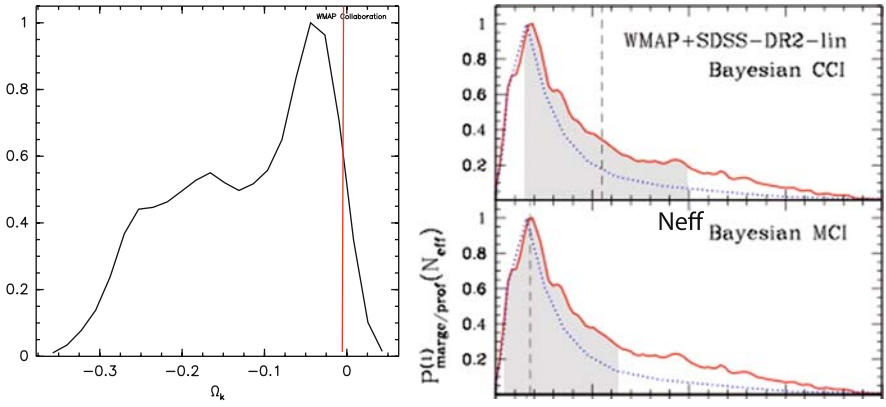

Fig. 3 Marginalization effects. *Left panel*: We consider the posterior distribution for the cosmological parameters of a dark energy + cold dark matter model where curvature is a free parameter and so is a (constant) equation of state parameter for dark energy. The data are the WMAP 5-year data. The *solid line* shows the N-dimensional maximum posterior value and the *black line* is the marginalized posterior over all other cosmological parameters. Figure courtesy of LAMBDA [5]. *Right panel*: figure from [6]. Illustration of central credible interval (CCI) and minimum credible interval (MCI), for the case of a ΛCDM model with free number of effective neutrino species (ignore *dotted line* for this example, *red line* is the marginalized posterior)

posterior volume lie in $(-\infty, \theta_{low})$ and (θ_{high}, ∞). This is called central credible interval and is connected to the median. Another possibility (minimum credible interval) is to consider the region so that the posterior at any point inside it is larger than at any point outside and so that the integral of the posterior in this region is the required fraction of the total. Thus remember, it is always a good practice to declare what confidence interval one is using. This subject is explored in more details in, e.g., [6].

7 Why Gaussian Likelihoods?

Throughout these lectures we always refer to Gaussian likelihoods. It is worth mentioning that if the data errors are Gaussianly distributed then the likelihood function for the data will be a multi-variate Gaussian. If the data are not Gaussianly distributed (but still are drawn from a distribution with finite variance!) we can resort to the central limit theorem: we can bin the data so that in each bin there is a superposition of many independent measurements. The central limit theorem will tell us that the resulting distribution (i.e., the error distribution for each bin) will be better approximated by a multi-variate Gaussian. However, as mentioned before, even if the data are Gaussianly distributed this does not ensure that the likelihood surface for the parameters will be a multi-variate Gaussian; for this to be always true the model needs to depend linearly on the parameters. Even without resorting to the central limit theorem, the Gaussian approximation is in many cases recovered even when starting from highly non-Gaussian distribution. A neat example is provided by Cash [7] which we follow here.

Let us say you want to constrain cosmology by studying cluster number counts as a function of redshift. The observation of a discrete number N of clusters is a Poisson process, the probability of which is given by the product

$$\mathcal{P} = \Pi_{i=1}^{N} [e_i^{n_i} \exp(-e_i)/n_i!], \tag{14}$$

where n_i is the number of clusters observed in the *i-th* experimental bin and e_i is the expected number in that bin in a given model, $e_i = I(x)\delta x_i$ with i being the proportional to the probability distribution. Here δx_i can represent an interval in clusters mass and/or redshift. Note: this is a product of Poisson distributions, thus one is assuming that these are independent processes. Clusters may be clustered, so when can this be used?

For unbinned data (or for small bins so that bins have only 0 and 1 counts) we define the quantity:

$$C \equiv -2 \ln \mathcal{P} = 2(E - \sum_{i=1}^{N} \ln I_i), \tag{15}$$

where E is the total expected number of clusters in a given model. The quantity ΔC between two models with different parameters has a χ^2 distribution! (so all that was said in Sect. 4 applies, even though we started from a highly non-Gaussian distribution.)

8 The Effect of Priors: Examples

Let us consider the two figures in Fig. 4. On the left, WMAP first-year data constraints in the Ω_m, Ω_Λ plane. On the right, models consistent with the WMAP 3-year data. In both cases the model is a non-flat ΛCDM model. So why the addition of more data (the two extra years of WMAP observations) gives worst constraints? The key is that what is reported in the plots is a representation of the posterior probability distribution. In the left panel a flat prior on Θ_A (angular size distance to the last scattering surface, giving by the position of the first peak) was assumed. In the figure on the right a flat prior on the Hubble constant H_0 was assumed. Remember: always declare the priors assumed!

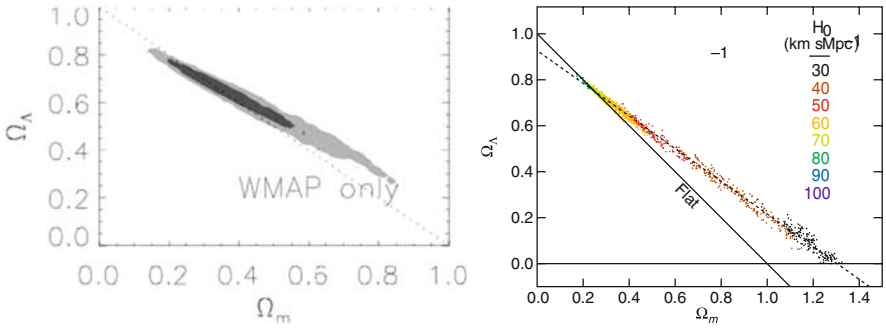

Fig. 4 *Left*: WMAP first-year data constraints in the Ω_m, Ω_Λ plane, from Spergel et al. [9]. *Right*: models consistent with the WMAP 3-year data, from Spergel et al. [8]. In both cases the model is a non-flat ΛCDM model. Figures reproduced with permission from the AAS

9 Combining Different Data Sets: Examples

It has become common to "combine data sets" and explore the constraints from the "data set combination." This means in practice that the likelihoods can be multiplied if the data sets are independent (if not the one should account for the appropriate covariance). It is important to note that *If the data-sets are inconsistent, the resulting constraints from the combined data set are nonsense*. An example is shown in Fig. 5.

On the left panel we show a figure from [8] constraints in the Ω_m, σ_8 plane for a flat ΛCDM model for WMAP 3-year data (blue), weak lensing constraints (orange), and combined constraints. On the right panel the figure shows the constraints in the

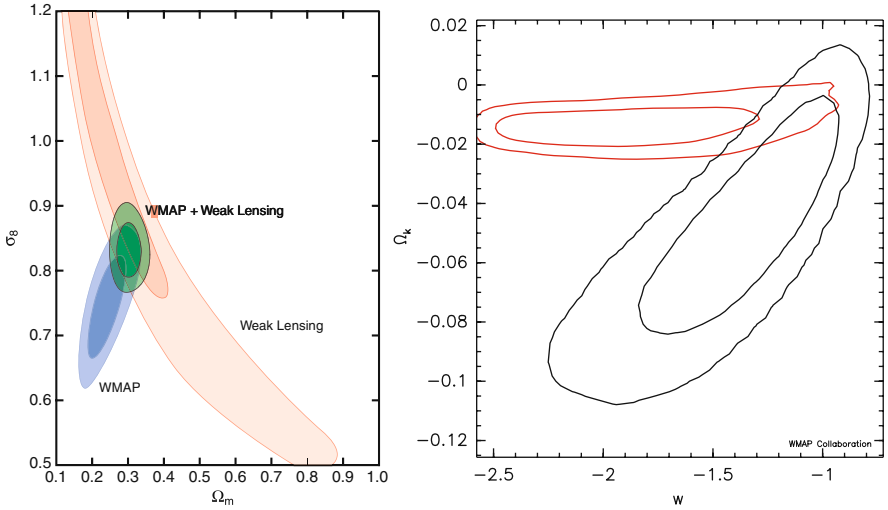

Fig. 5 *Left*: constraints in the Ω_m, σ_8 plane for a flat ΛCDM model for WMAP 3-year data, weak lensing constraints, and combined constraints. Figure from Spergel et al. [8], reproduced with permission from the AAS. *Right*: Constraints in the Ω_k,w plane for non-flat dark energy models with constant w for WMAP5+supernovae data (*lower curves*) and WMAP5+BAO (*upper curves*). Figure courtesy of LAMBDA [5]

Ω_k,w plane for non-flat dark energy models with constant w for WMAP5+ supernovae data (in black) and WMAP5+BAO (in red). Even though the WMAP data are in common there is some tension in the resulting constraints. The two data sets (supernovae and BAO, WMAP and weak lensing) are not fully consistent, as the authors themselves, note, they should not be combined.

10 Forecasts: Fisher Matrix

Before diving into the details let us re-examine the error estimates for parameters from the likelihood. Let us assume a flat prior in the parameter so we can identify the posterior with the likelihood. Close to the peaks we can expand the log likelihood in Taylor series

$$\ln \mathcal{L} = \ln \mathcal{L}(\theta_0) + \frac{1}{2} \sum_{ij} (\theta_i - \theta_{i,0}) \left. \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \right|_{\theta_0} (\theta_j - \theta_{j0}) + \ldots . \tag{16}$$

By truncating this expansion to the quadratic term (remember that by expanding around the maximum we have the first derivative equal to zero) we say the likelihood surface is locally a multi-variate Gaussian. The Hessian matrix is defined as

$$\mathcal{H}_{ij} = -\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j}. \tag{17}$$

It encloses the information of the parameters' errors and their covariance. If this matrix is not diagonal it means that the parameters' estimates are correlated. Loosely speaking we said "the parameters are correlated": it means that they have a similar effect on the data and thus the data have hard time in telling them apart. The parameters may or may not be physically related with each other.

More specifically if all parameters are kept fixed except one (parameter i, say), the error on that parameter would be given by $1/\sqrt{\mathcal{H}_{ii}}$. This is called conditional error but is almost never used or interesting.

Having understood this, we can move on to the Fisher information matrix [10]. The Fisher matrix plays a fundamental role in forecasting errors from a given experimental set up and thus is the work-horse of experimental design. It is defined as

$$F_{ij} = -\left\langle \frac{\partial^2 \ln \mathcal{L}}{\partial\theta_i \partial\theta_j} \right\rangle. \tag{18}$$

It should be clear that $F = \langle \mathcal{H} \rangle$.

Here the average is the ensemble average over observational data (those that would be gathered if the real Universe was given by the model – and model parameters – around which the derivative is taken). As we have seen the likelihood for independent data sets is the product of the likelihoods, it follows that the Fisher matrix for independent data sets is the sum of the individual Fisher matrices. This will become useful later on.

In the one-parameter case, say only i component of θ, thinking back at the Taylor expansion around the maximum of the likelihood we have that

$$\Delta \ln \mathcal{L} = \frac{1}{2} F_{ii} (\theta_i - \hat{\theta}_i)^2 \tag{19}$$

when $2\Delta \ln \mathcal{L} = 1$ and by identifying it with the $\Delta\chi^2$ corresponding to 68% confidence level, we see that $1/\sqrt{F_{ii}}$ yields the $1-\sigma$ displacement for θ_i. This is the analogous to the conditional error from above. In the general case

$$\sigma_{ij}^2 \geq (F^{-1})_{ij}. \tag{20}$$

Thus when all parameters are estimated simultaneously from the data the marginalized error is

$$\sigma_{\theta_i} \geq (F^{-1})_{ii}^{1/2}. \tag{21}$$

Let's spell it out for clarity: this is the square root of the element ii of the inverse of the Fisher information matrix.[2] This assumes that the likelihood is a Gaussian

[2] i.e., you have to perform a matrix inversion first.

around its maximum (the fact that the data are Gaussianly distributed is no guarantee that the likelihood will be Gaussian, see, e.g., Fig. 2). The terrific utility of the Fisher Information matrix is that, if you can compute it, it enables you to estimate the parameters errors *before you do the experiment*. If it can be computed quickly, it also enables one to explore different experimental setups and optimize the experiment. This is why the Fisher matrix approach is so useful in survey design. Also complementarity of different, independent, and uncorrelated experiments (i.e., how in combination they can lift degeneracies) can be quickly explored: the combined Fisher matrix is the sum of the individual matrices. This is of course extremely useful; however, read below for some caveats.

The $\geq$ is the Kramer–Rao inequality: the Fisher matrix approach always gives you an optimistic estimate of the errors (reality is only going to be worst). And this is not only because systematic and real-world effects are often ignored in the Fisher information matrix calculation, but for a fundamental limitation: only if the likelihood is Gaussian that $\geq$ becomes $=$. In some cases, when the Gaussian approximation for the likelihood does not hold, it is possible to make non-linear transformation of the parameter that makes the likelihood Gaussian. Basically, if the data are Gaussianly distributed and the model depends linearly on the parameters then the likelihood would be Gaussian. So the key is to have a good enough understanding of the theoretical model to be able to find such a transformation. See [11] for a clear example.

10.1 Computing Fisher Matrices

The simplest, brute force approach to compute a Fisher matrix is as follows: write down the likelihood for the data given the model. Instead of the data values (which are not known) use the theory prediction for a fiducial model. This will add a constant term to the log likelihood which does not depend on cosmology. In the covariance matrix include expected experimental errors. Then take derivatives with respect to the parameters as indicated in (18).

In the case where the data are Gaussianly distributed it is possible to compute explicitly and analytically the Fisher matrix, in a much more elegant way than above:

$$F_{ij} = \frac{1}{2} Tr\left[C^{-1}C_{,i}C^{-1}C_{,j} + C^{-1}M_{ij}\right], \tag{22}$$

where $M_{ij} = y_{,i}y_{j}^{T} + y_{,j}y_{,i}^{T}$ and $,i$ denotes derivative with respect to the parameter θ_i. This is extremely useful, you need to know the covariance matrix (which may depend on the model and need not be diagonal) and you need to have a fiducial model y which you know how it depends on the parameter θ. Then the Fisher matrix gives you the expected (forecasted) errors. Priors or forecasts results from other experiments can be easily included by simply adding their Fisher before

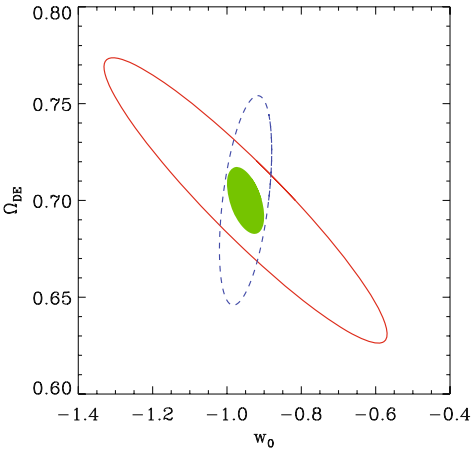

Fig. 6 Marginalized 68% CL constraints on the dark energy parameters expected for the DUNE weak lensing (*dashed*), a full sky BAO survey (*solid*), and their combination (*solid filled*). This figure was derived using the Fisher matrix routines of iCosmo. Figure from Refregier et al. [12]

performing the matrix inversion to obtain the marginal errors. This is illustrated in Fig. 6, from [12] and produced using the icosmo (http://www.icosmo.org/) software.

Before we finish this section let us spell out the following prescription.

Imagine you have computed a large Fisher matrix, varying all parameters Ω_k, w_0, neutrino mass m_ν, number of neutrino species N_ν, running of the spectral index α, etc. Now you want to compute constraints for a standard flat ΛCDM model. Simply ignore row and columns corresponding to the parameters that you want to keep fixed at the fiducial value before inverting the matrix.

Imagine now that you have a six parameters' Fisher matrix (say H_0, Ω_m,τ, Ω_Λ, n, Ω_b, σ_8), and want to produce 2D plots for the confidence regions for parameters 2 and 4, say, marginalized over all other (1,3,5,6) parameters. Invert F_{ij}. Take the sub-matrix made by rows and columns corresponding to the parameters of interest (2 and 4 in this case) and invert back this sub-matrix.

The resulting matrix, let us call it $\mathcal{Q}$, describes a Gaussian 2D likelihood surface in the parameters 2 and 4 or, in other words, the chisquare surface for parameters 2,4 – marginalized over all other parameters – can be described by the equation

$$\widetilde{\chi}^2 = \sum_{ij} (\theta_i - \theta_i^{fid.}) \mathcal{Q}_{ij} (\theta_j - \theta_j^{fid.}) \,. \tag{23}$$

From this equation, getting the errors corresponds to finding the quadratic equation solution $\widetilde{\chi}^2 = \Delta\chi^2$. For correspondence between $\Delta\chi^2$ and confidence region see the earlier discussion. If you want to make plots, the equation for the elliptical boundary for the joint confidence region in the sub-space of parameters of interest is $\Delta = \delta\theta\,\mathcal{Q}^{-1}\delta\theta$.

11 Example of Fisher Approach Applications

Here we are going to consider two cases of application of Fisher forecasts that are extensively used in the literature. This section assumes that the reader is familiar with basic CMB and large-scale structure concepts, such as power spectra, error on power spectra, cosmic variance, window and selection function, instrumental noise and shot noise, redshift space. Some readers may find this section more technical than the rest of this document; it is possible to skip it and continue reading from Sect. 12.

11.1 CMB

The CMB has become the single data set that gives most constraints on cosmology. As the recently launched Planck satellite will yield the ultimate survey for primary CMB temperature anisotropies, doing Fisher matrix forecasts of CMB temperature data may very soon be obsolete. There remains the scope for forecasting constraints from polarization experiments, however, systematic effects (e.g., foreground subtraction) will likely dominate the statistical errors (see, e.g., [13] for details). It is still, however, a good exercise to see how one can set up a Fisher matrix analysis for CMB data.

If we have a noiseless full sky survey and the initial conditions are Gaussian we can write that the signal in the sky (i.e., the spherical harmonic transform of the anisotropies) is Gaussianly distributed. We can write the signal as

$$\mathbf{s}_\ell = (a_\ell^T, a_\ell^E, a_\ell^B), \tag{24}$$

where a_{ell} denotes the spherical harmonic coefficients for temperature and E and B model polarization. The covariance matrix $\mathbf{C}_\ell$ is then given by

$$\mathbf{C}_\ell = \begin{pmatrix} C_\ell^{TT} & C_\ell^{TE} & 0 \\ C_\ell^{TE} & C_\ell^{EE} & 0 \\ 0 & 0 & C_\ell^{BB} \end{pmatrix}, \tag{25}$$

where C_ℓ denotes the angular CMB power spectrum. Using (22) and considering that, for rotational invariance, for every ℓ there are $(2\ell + 1)$ modes, it is possible to show that the Fisher matrix for CMB experiments can be rewritten as

$$F_{ij}^{CMB} = \sum_{XY} \sum_{\ell} \frac{\partial C_\ell^X}{\partial \theta_i} \left(C_\ell^{XY}\right)^{-1} \frac{\partial C_\ell^Y}{\partial \theta_j}, \tag{26}$$

where for the matrix $\mathcal{C}_\ell$ the elements are $\mathcal{C}_\ell^{XY}$, where X,Y=TT, TE, EE, BB, etc., is given by[3]

$$\mathcal{C}_\ell = \frac{2}{2\ell+1}\begin{pmatrix} (C_\ell^{TT})^2 & (C_\ell^{TE})^2 & C_\ell^{TT}C_\ell^{TE} & 0 \\ (C_\ell^{TE})^2 & (C_\ell^{EE})^2 & C_\ell^{EE}C_\ell^{TE} & 0 \\ C_\ell^{TT}C_\ell^{TE} & C_\ell^{EE}C_\ell^{TE} & 1/2[(C_\ell^{TE})^2 + C_\ell^{TT}C_\ell^{EE}] & 0 \\ 0 & 0 & 0 & (C_\ell^{BB})^2 \end{pmatrix}. \quad (27)$$

Note that this matrix is more complicated that what one would have obtained by assuming a Gaussian distribution for the C_ℓ and no correlation between TT, TE, and EE. Nevertheless, (26) is simple enough and allows one to quickly compute forecasts from ideal CMB experiments.

In this formalism effects of partial sky coverage and of instrumental noise can be included (approximatively) by the following substitutions:

$$C_\ell \longrightarrow C_\ell + N_\ell, \quad (28)$$

where N_ℓ denotes the effective noise power spectrum. Note that N_ℓ depends on ℓ even for a perfectly white noise because of beam effects. In addition the partial sky coverage can be accounted for by considering that the number of independent modes decreases with the sky coverage: if f_{sky} denotes the fraction of sky covered by the experiment, then

$$\mathcal{C}_\ell \longrightarrow \mathcal{C}_\ell / f_{sky}. \quad (29)$$

11.2 Baryon Acoustic Oscillations

Cosmological perturbations in the early Universe excite sound waves in the photon–baryon fluid. After recombination, these baryon acoustic oscillations (BAO) became frozen into the distribution of matter in the Universe imprinting a preferred scale, the sound horizon. This defines a standard ruler whose length is the distance sound can travel between the Big Bang and recombination. The BAO are directly observed in the CMB angular power spectrum and have been observed in the spatial distribution of galaxies by the 2dF GRS survey and the SDSS survey [14–16]. The BAO, observed at different cosmic epochs, act as a powerful measurement tool to probe the expansion of the Universe, which in turn is a crucial handle to constrain the nature of dark energy. The underlying physics which sets the sound horizon scale ($\sim$150 Mpc comoving) is well understood and involves only linear perturbations in the early Universe. The BAO scale is measured in surveys of galaxies from the statistics of the three-dimensional galaxy positions. Only recently have galaxy

[3] I owe this proof to P. Adshead.

surveys such as SDSS grown large enough to allow for this detection. The existence of this natural standard measuring rod allows us to probe the expansion of the Universe. The angular size of the oscillations in the CMB revealed that the Universe is close to flat. Measurement of the change of apparent acoustic scale in a statistical distribution of galaxies over a large range of redshift can provide stringent new constraints on the nature of dark energy. The acoustic scale depends on the sound speed and the propagation time. These depend on the matter to radiation ratio and the baryon-to-photon ratio. CMB anisotropy measures these and hence fixes the oscillation scale. A BAO survey measures the acoustic scale along and across the line of sight. At each redshift, the measured angular (transverse) size of oscillations, $\Delta\theta$, corresponds with the physical size of the sound horizon, where the angular diameter distance D_A is an integral over the inverse of the evolving Hubble parameter, $H(z)$. $r_\perp = (1+z)D_A(z)\delta\theta$. In the radial direction, the BAO directly measure the instantaneous expansion rate $H(z)$, through $r_\parallel = (c/H(z))\Delta z$, where the redshift interval (Δz) between the peaks is the oscillation scale in the radial direction. As the true scales $r_\perp$ and $r_\parallel$ are known (given by r_s, the sound horizon at radiation drag, well measured by the CMB) this is not an Alcock–Paczynski test but a "standard ruler" test. Note that in this standard ruler test the cosmological feature used as the ruler is not an actual object but a statistical property: a feature in the galaxy correlation function (or power spectrum). An unprecedented experimental effort is undergoing to obtain galaxy surveys that are deep, larger, and accurate enough to trace the BAO feature as a function of redshft. Before these surveys can even be designed it is crucial to know how well a survey with given characteristic will do. This was illustrated very clearly in [17], which we follow closely here. We will adopt the Fisher matrix approach. To start we need to compute the statistical error associated to a determination of the galaxy power spectrum $P(k)$. In what follows we will ignore the effects of non-linearities and complicated biasing between galaxies and dark matter: we will assume that galaxies, at least on large scales, trace the linear matter power spectrum in such a way that their power spectrum is directly proportional to the dark matter one: $P(k) = b^2 P_{DM}(k)$ where b stands for galaxy bias. At a given wavevector k, the statistical error of the power spectrum is a sum of a cosmic variance term and a shot noise term:

$$\frac{\sigma_P(k)}{P(k)} = \frac{P(k) + 1/n}{P(k)}. \tag{30}$$

Here n denotes the average density of galaxies and $1/N$ is the white noise contribution from the fact that galaxies are assumed to be a Poisson sampling of the underlying distribution. When written in this way this expression assumes that n is constant with position. While in reality this is not true for forecasts, one assumes that the survey can be divided in shells in redshifts and that the selection function is such that n is constant within a given shell. Since $P(k)$ is also expected to change in redshift then one should really implicitly assume that there is a z dependence in (30). In general $P(k,z) = b(z)^2 G^2(z) P_{DM}(k)$ where $G(z)$ denotes the linear growth factor, i.e., the bias is expected to evolve with redshift as well as clustering does, not only

because galaxy bias changes with redshift but also because at different redshifts one may be seeing different type of galaxies which may have different bias parameter. We do not know a priori the form of $b(z)$ but given a fiducial cosmological model we know $G(z)$. Preliminary observations seem to indicate that the z evolution of b tends to cancel that of $G(z)$, so it is customary to assume that $b(z)G(z) \sim$ constant, but we should bear in mind that this is an assumption.

An extra complication arises because galaxy redshift surveys use the redshift as distance indicator, and deviations from the Hubble flow therefore distort the clustering. If the Universe was perfectly uniform and galaxies were test particles these deviations from the Hubble flow would not exist and the survey would not be distorted. But clustering does perturb the Hubble for and thus introduces the so-called redshift-space distortions in the clustering measured by galaxy redshift surveys. Note that redshift-space distortions only affect the line-of-sight clustering (it is a perturbation to the distances) not the angular clustering. Since these distortions are created by clustering they carry, in principle, important cosmological information. To write this dependence explicitly

$$P(k,\mu,z) = b(z)^2 G(z)^2 P_{DM}(k)(1+\beta\mu)^2, \tag{31}$$

where μ denotes the cosine of the angle between the line-of-sight and the wavevector $\beta = f/b = d\ln G(z)/d\ln a/b \simeq \Omega_m(z)^{0.6}/b$. In the linear regime, the cosmological information carried by the redshift-space distortions is enclosed in the $f(z) = \beta(z)b(z)$ combination.

For finite surveys, $P(k)$ at nearby wavenumbers are highly correlated, the correlation length is related to the size of the survey volume; for large volumes the cell size over which modes are correlated is $(2\pi)^3/V$ where V denotes the comoving survey volume. Only over distances in k-space larger than that modes can be considered independent. If one therefore wants to count over all the modes anyway (for example, by transforming discrete sums into integrals in the limit of large volumes) then each k needs to be downweighted, to account the fact that all k are not independent. In addition one should keep in mind that Fourier modes $\mathbf{k}$ and $-\mathbf{k}$ are not independent (the density field is real valued!), giving an extra factor of 2 in the weighings. We can thus write the error on a band power centered around k,

$$\frac{\sigma_P}{P} = 2\pi\sqrt{\frac{2}{Vk^2\delta k\Delta\mu}}\left(\frac{1+nP}{nP}\right). \tag{32}$$

In the spirit of the Fisher approach we now assume that the likelihood function for the band powers $P(k)$ is Gaussian, thus we can approximate the Fisher matrix by

$$F_{ij} = \int_{k_{\rm min}}^{k_{\rm max}} \frac{\partial \ln P(\mathbf{k})}{\partial\theta_i}\frac{\partial \ln P(\mathbf{k})}{\partial\theta_j} V_{\rm eff}(\mathbf{k})\frac{d\mathbf{k}}{2(2\pi)^3}. \tag{33}$$

The derivatives should be evaluated at the fiducial model and $V_{\rm eff}$ denotes the effective survey volume given by

$$V_{\rm eff}(\mathbf{k}) = V_{\rm eff}(k,\mu) = \int \left[\frac{n(z)P(k,\mu)}{n(z)P(k,\mu)+1}\right]^2 dz = \left[\frac{nP(k,\mu)}{nP(k,\mu)+1}\right]^2 V, \qquad (34)$$

where $n = \langle n(z)\rangle$. Equation (33) can be written explicitely as a function of k and μ as

$$F_{ij} = \int_{-1}^{1}\int_{k_{\rm min}}^{k_{\rm max}} = \frac{\partial \ln P(k,\mu)}{\partial\theta_i}\frac{\partial \ln P(k,\mu)}{\partial\theta_j} V_{\rm eff}(k,\mu)\frac{k^2 dk d\mu}{2(2\pi)^2}\,. \qquad (35)$$

In writing this equation we have assumed that over the entire survey extension the line-of-sight direction does not change: in other words, we made the flat sky approximation. For forecasts this encloses all the statistical information anyway, but for actual data analysis application the flat sky approximation may not hold. In this equation $k_{\rm min}$ is set by the survey volume; for future surveys where the survey volume is large enough to sample the first BAO wiggling the exact value of $k_{\rm min}$ does not matter, however, recall that for surveys of typical size L (where $L \sim V^{1/3}$), the largest scale probed by the survey will be corresponding to $k = 2\pi/L$. Keeping in mind that the first BAO wiggle happens at ~ 150 Mpc the survey size needs to be $L \gg 150$ Mpc for $k_{\rm min}$ to be unimportant and for the "large volume approximation" made here to hold. As anticipated above, one may want to sub-divide the survey into independent redshift shells, compute the Fisher matrix for each shell, and then combine the constraints. In this case L will be set by the smallest dimension of the volume (typically the width of the shell), so one needs to make sure that the width of the shell still guarantees a large volume and large L. $k_{\rm max}$ denotes the maximum wavevector to use. One could, for example, impose a sharp cut to delimit the range of validity of linear theory. In [18] this is improved as we will see below.

Before we do that, let us note that there are two ways to interpret the parameters θ_{ij} in (35). One could simply assume a cosmological model, say, for example, a flat quintessence model where the equation of state parameter $w(z)$ is parameterized by $w(z) = w(0) + w_a(1-a)$ and take derivatives of $P(k,\mu)$ with respect to these parameters. Alternatively, one could simply use as parameters the quantities $H(z_i)$ and $D_A(z_i)$, where z_i denotes the survey redshift bins. These are the quantities that govern the BAO location and are more general; they allow one not to choose a particular dark energy model until the very end. Then one must also consider the cosmological parameters that govern the $P(k)$ shape $\Omega_m h^2$, $\Omega_b h^2$, and n_s. Of course one can also consider $G(z_i)$ as free parameters and constrain these either through the overall $P(k)$ amplitude (although one would have to assume that $b(z)$ is known, which is dicey) or through the determination of $G(z)$ and $\beta(z)$. The safest and most conservative approach, however, is to ignore any possible information coming from $G(z)$, $\beta(z)$, or n_s and to only try to constrain expansion history parameters.

The piece of information still needed is how the expansion history information is extracted from $P(k,\mu)$. When one converts **ra, dec**, and redshifts into distances and positions of galaxies of a redshift survey, one assumes a particular reference cosmology. If the reference cosmology differs from the true underlying cosmology, the inferred distances will be wrong and so the observed power spectrum will be distorted:

$$P(k_\perp, k_\parallel) = \frac{D_a(z)^2_{ref} H(z)_{\rm true}}{D_A(z)^2_{\rm true} H(z)_{\rm ref}} P_{\rm true}(k_\perp, k_\parallel). \tag{36}$$

Note that since distances are affected by the choice of cosmology and k vectors are $k_{{\rm ref},\parallel} = H(z)_{\rm ref}/H(z)_{\rm true} k_{{\rm true},\parallel}$ and $k_{{\rm ref},\perp} = D_A(z)_{\rm true}/D_A(z)_{\rm ref} k_{{\rm true},\perp}$. Note that therefore in (36) we can write

$$P_{\rm true}(k_\perp, k_\parallel, z) = b(z)^2 \left(1 + \beta(z) \frac{k^2_{{\rm true},\parallel}}{k^2_{{\rm true},\perp} + k^2_{{\rm true},\parallel}}\right)^2 \left[\frac{G)(z)}{G(z_o)}\right]^2 P_{DM}(k, z_o), \tag{37}$$

where z_o is some reference redshift where to normalize $P(k)$ typical choices can be the CMB redshift or redshift $z = 0$. Not that from these equations it should be clear that what the BAO actually measure directly is $H(z)r_s$ and D_A/r_s where r_s is the BAO scale, the advantage is that r_s is determined exquisitely from the CMB.

How would then one convert these constraints on those on a model parameter? Clearly, one then projects the resulting Fisher matrix on the dark energy parameters space. In general if you have a set of parameters θ_i with respect to which the Fisher matrix has been computed, but you would like to have the Fisher matrix for a different set of parameters ϕ_i, where the θ_i are functions of the ϕ_i, the operation to implement is

$$F_{\phi_i,\phi_j} = \sum_{mn} \frac{\partial \theta_n}{\partial \phi_i} F_{\theta_n,\theta_m} \frac{\partial \theta_m}{\partial \phi_j}. \tag{38}$$

The full procedure for the BAO survey case is illustrated in Fig. 7. The slight complication is that one starts off with a Fisher matrix (for the original parameter set θ_i) where some parameters are nuisance and need to be marginalized over, so some matrix inversions are needed.

So far non-linearities have been just ignored. It is, however, possible to include then at some level in this description. Reference [18] proceed by introducing a distribution of Gaussianly distributed random displacements parallel or perpendicular to the line-of-sight coming from non-linear growth (in all directions) and from non-linear redshift-space distortions (only in the radial direction). The publicly available code that implements all this (and more) is at http://cmb.as.arizona.edu/ eisenste/ acousticpeak/bao_forecast.html. In order to use the code keep in mind that in Ref. [18] the authors model the effect of non-linearities by convolving the galaxy distribution with a redshift dependent and μ-dependent smoothing kernel. The

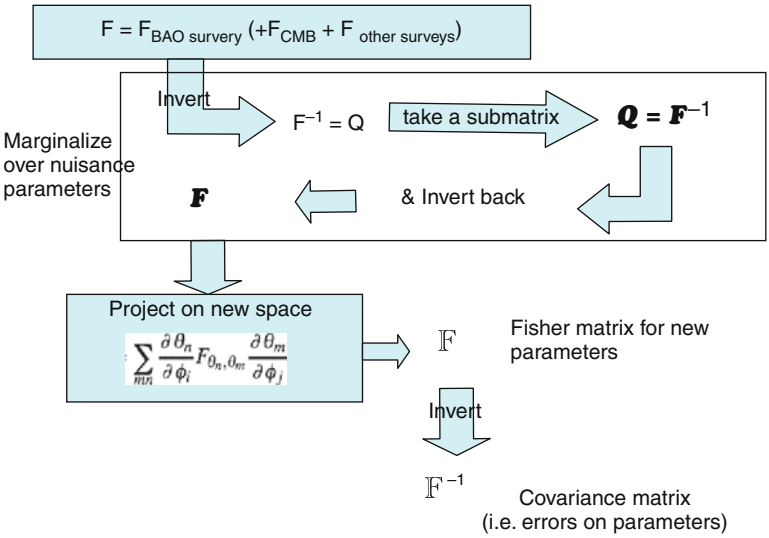

Fig. 7 Steps to implement once the Fisher matrix of (35) has been computed to obtain error on dark energy parameters

effect on the power spectrum is to multiply $P(k)$ by $\exp[-k^2\Sigma(k,\mu)/2]$, where $\Sigma(k,\mu) = \Sigma_\perp^2 - \mu^2(\Sigma_\parallel^2 - \Sigma_\perp^2)$. As a consequence the integrand of the Fisher matrix expression of (35) is multiplied by

$$\exp[-k^2\Sigma_\perp^2 - k^2\mu^2(\Sigma_\parallel^2 - \Sigma_\perp^2)], \tag{39}$$

where, to be conservative, the exponential factor has been taken outside the derivatives, which is equivalent to marginalize over the parameters $\Sigma_\parallel$ and $\Sigma_\perp$ with large uncertainties.

Note that $\Sigma_\parallel$ and $\Sigma_\perp$ depend on redshift and on the chosen normalization for $P_{DM}(k)$. In particular,

$$\Sigma_\perp(z) = \Sigma_0 G(z)/G(z_0), \tag{40}$$

$$\Sigma_\parallel(z) = \Sigma_0 G(z)/G(z_0)(1+f(z)), \tag{41}$$

$$\Sigma_0 \propto \sigma_8\,. \tag{42}$$

If in your convention $z_0 = 0$ then $\Sigma_0(z=0) = 8.6h^{-1}\sigma_{8,DM}(z=0)/0.8$.

As an example of an application of this approach for survey design, it may be interesting to ask the question of what is the optimal galaxy number density for a given survey. Taking redshifts is expensive and for a given telescope time allocated, only a certain number of redshifts can be observed. Thus is it better to survey more volume but have a low number density or survey a smaller volume with higher number density? You can try to address this issue using the available code. For a cross check, Fig. 8 shows what you should obtain. Here we have assumed $\sigma_8 = 0.8$

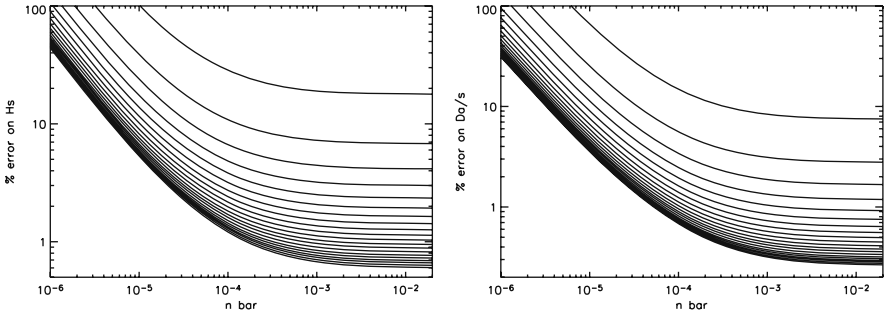

Fig. 8 Percent error on $H(z)r_s$ and D_a/r_s as a function of the galaxy number density of a BAO survey. This figure assumes full sky coverage $f_{sky} = 1$ (errors will scale like $1/\sqrt{f_{\rm sky}}$) and redshift range from $z = 0$ to $z = 2$ in bins of $\Delta z = 0.1$

at $z = 0$, $b(z = 0) = 1.5$ and we have assumed that $G(z)b(z) =$ constant. To interpret this figure note that with the chosen normalizations, $P(k)$ in real space at the BAO scale $k \sim 0.15$ h/Mpc is 6241(Mpc/h)^3, boosted up by large-scale redshift-space distortions to roughly 10^4(Mpc/h)^3 so $n = 10^{-4}$ corresponds to $nP(k = 0.15) = 1$. Note that the "knee" in this figure is therefore around $nP = 1$. This is where this "magic number" of reaching $nP \sim> 1$ in a survey comes from. Of course, there are other considerations that would tend to yield an optimal nP bigger than unity and of the order of few.

12 Model Testing

So far we have assumed a cosmological model characterized by a given set of cosmological parameters and used statistical tools to determine the best fit for these parameters and confidence intervals. However, the best fit parameters and confidence intervals depend on the underlying model, i.e., what set of parameters are allowed to vary. For example, the estimated value for the density parameter of baryonic matter Ω_b changes depending whether in a ΛCDM model the Universe is assumed flat or not (Fig. 9 right panel) or the recovered value for the spectral slope of the primordial power spectrum changed depending if the primordial power spectrum is assumed to be a power law or is allowed to have some "curvature" or "running" (Fig. 9 left panel). It would be useful to be able to allow the data to determine which combination of parameters gives the preferred fit to the data; this is the problem of *model selection*. Here we start by following [19] which is a clear introduction to the application of this subject in cosmology. Model selection relies on the so-called information criteria and the goal is to make an objective comparison of different *models* which may have a different number of parameters. The models considered in the example of Fig. 9 are "nested" as one model (the ΛCDM one) is completely specified by a sub-set of the parameters of the other

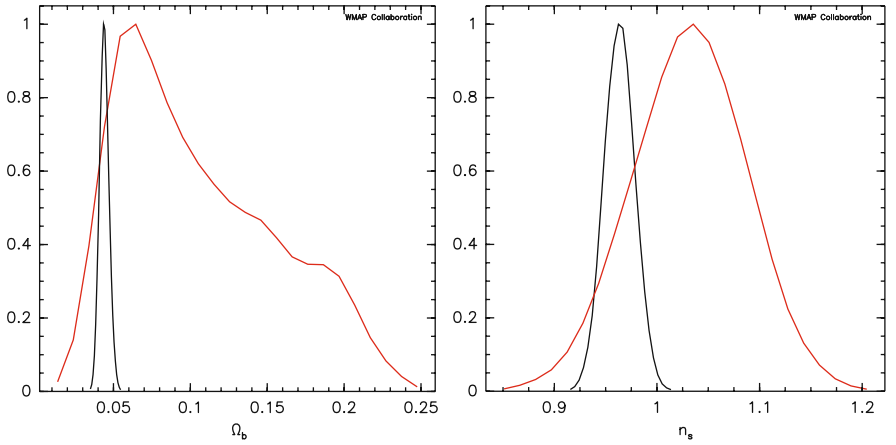

Fig. 9 Effect of the choice of the cosmological model in the recovered values for the parameters. Here we used WMAP5 data only: in both panels the *narrow curve* is for a standard flat ΛCDM model. In the *left panel* we show the posterior for Ω_b, the *broad curve* is for a non-flat ΛCDM model. In the *right panel* we show the posterior for n_s: the *broad curve* is for a ΛCDM model where the primordial power spectrum is not a perfect power law but is allowed to have some "curvature" also called "running" of the spectral index. Figure courtesy of LAMBDA [5]

(more general) model. In cosmology one is almost always concerned with nested models.

Typically the introduction of extra parameters will yield an improved fit to the data set, so a simple comparison of the maximum likelihood value will always favor the model with more parameters, regardless of whether the extra parameters are relevant. There are several different approaches often used in the literature. The simplest is the likelihood ratio test [20] see Sect. 6. Consider the quantity $2\ln[\mathcal{L}_{\rm simple}/\mathcal{L}_{\rm complex}]$ where $\mathcal{L}_{\rm simple}$ denotes the maximum likelihood for the model with less parameters and $\mathcal{L}_{\rm complex}$ the maximum likelihood for the other model. This quantity is approximately chisquare distributed and thus the considerations of Sect. 4 can be applied.

The Akaike information criterion (AIC) [21] is defined as $\mathrm{AIC} = -2\ln\mathcal{L} + 2k$ where $\mathcal{L}$ denotes the maximum likelihood for the model and k the number of parameters of the model. The best model is the one that minimizes AIC.

The Bayesian information criterion (BIC)[22] is defined as $\mathrm{BIC} = -2\ln\mathcal{L} + k\ln N$ where N is the number of data points used in the fit.

It should be clear that all these approaches tend to downweigh the improvement in the likelihood value for the more complex model with a penalty that depends on how complex is the model. Each of these approaches has its pros and cons and there is no silver bullet.

However, it is possible to place model selection on firm statistical grounds within the Bayesian approach by using the *Bayesian factor* which is the Bayesian evidence ratio (i.e., the ratio of probabilities of the data given the two models).

Recalling the Bayes theorem (2) we can write $\mathcal{P}(D) = \sum_i \mathcal{P}(D|M_i)\mathcal{P}(M_i)$ where i runs over the models M we are considering. Then the Bayesian evidence is

$$\mathcal{P}(D|M_i) = \int d\theta \mathcal{P}(D|\theta,M_i)\mathcal{P}(\theta|M_i), \tag{43}$$

where $\mathcal{P}(D|\theta,M_i)$ is the likelihood. Given two models (i and j), the Bayes factor is

$$B_{ij} = \frac{\mathcal{P}(D|M_i)}{\mathcal{P}(D|M_j)}. \tag{44}$$

A large B_{ij} denotes preference for model i. In general this requires complex numerical calculations, but for the simple case of Gaussian likelihoods it can be expressed analytically. The details can be found, e.g., in [23] and references therein. For a didactical introduction see also [24].

13 Monte Carlo Methods

With the recent increase in computing power, in cosmology we resort to the application of Monte Carlo methods ever more often. There are two main applications of Monte Carlo methods: Monte Carlo error estimations and Markov Chains Monte Carlo. Here I will concentrate on the first as there are several basics and detail explanations of the second (see e.g., [25] and references therein).

Let us go back to the issue of parameter estimation and error calculation. Here is the conceptual interpretation of what it means that an experiment measures some parameters (say cosmological parameters). There is some underlying true set of parameters $\theta_{\rm true}$ that are only known to Mother Nature but not to the experimenter. There true parameters are statistically realized in the observable Universe and random measurement errors are then included when the observable Universe gets measured. This realization gives the measured data D_0. Only D_0 is accessible to the observer (you). Then you go and do what you have to do to estimate the parameters and their errors (chisquare, likelihood, etc.) and get θ_0. Note that D_0 is not a unique realization of the true model given by $\theta_{\rm true}$: there could be infinitely many other realizations as hypothetical data sets, which could have been the measured one: D_2,D_2,D_3 ... each of them with a slightly different fitted parameters θ_1, θ_2 ... θ_0 is one parameter set drawn from this distribution. The hypothetical ensemble of Universes described by θ_i is called ensemble, and one expects that the expectation value $\langle\theta_i\rangle = \theta_{\rm true}$. If we knew the distribution of $\theta_i - \theta_{\rm true}$ we would know everything we need about the uncertainties in our measurement θ_0. The goal is to infer the distribution of $\theta_i - \theta_{\rm true}$ without knowing $\theta_{\rm true}$. Here is what we do we say that hopefully θ_0 is not too wrong and we consider a fictitious world where θ_0 was the true one. So it would not be such a big mistake to take the probability distribution of $\theta_i - \theta_0$ to be that of $\theta_i - \theta_{\rm true}$. In many cases we know how to simulate $\theta_i - \theta_0$ and so we can simulate many synthetic realization of "worlds"

where θ_0 is the true underlying model. Then mimic the observation process of these fictitious Universes replicating all the observational errors and effects and from each of these fictitious Universe estimate the parameters. Simulate enough of them and from $\theta_i^S - \theta_0$ (where S stands for "synthetic" or "simulated") you will be able to map the desired multi-dimensional probability distribution. With the advent of fast computers this technique has become increasingly widespread. As long as you believe you know the underlying distribution and that you believe you can mimic the observation replicating all the observational effects this technique is extremely powerful and, I would say, indispensable. This is especially crucial when complicated effects such as instrumental and or systematic effects can be simulated but not described analytically by a model.

14 Conclusions

I have given a brief overview of statistical techniques that are frequently used in the cosmological literature. I have presented several examples often from the literature to put these techniques into context. This is not an exhaustive list nor a rigorous treatment, but a starter kit to "get you started." As more and more sophisticated statistical techniques are used to make the most of the data, one should always remember that they need to be implemented and used correctly:

- data gathering is an expensive and hard task; statistical techniques make possible to make the most of the data
- always beware of systematic effects
- an incorrect treatment of the data will give non-sensical results
- there will always be things that are beyond the statistical power of a given data set

Remember: "Treat your data with respect!"

15 Some Useful References

There are many good and rigorous statistics books out there. In particular Kendall's advanced theory of statistics made of three volumes are

- *Distribution Theory* (Stuart and Ort 1994 [26])
- *Classical Inference* (Stuart and Ort 1991 [27]) and
- *Bayesian Inference* (O'Hagan 1994 [20]).

For astronomical and cosmological applications in many cases one may need a practical manual rather than a rigorous textbook. Although it is important to note that a practical manual is no substitute for a rigorous introduction to the subject.

- *Practical Statistics for Astronomers*, by Wall and Jenkins, (2003) is a must have [1].
- *Numerical Recipes* is also an indispensable "bible": Press et al. [2]

It also provides a guide to the numerical implementation of the "recipes" discussed. Complementary information to what presented here can be found in

- Verde, in XIX Canary Island Winter School "*The Cosmic Microwave Background: From Quantum Fluctuations to the Present Universe*" [25]. In the form of lecture notes, and
- Martinez, Saar, "*Statistics of the Galaxy Distribution*" [28], with a slant on large scale structure and data analyis in cosmology, Martinez, Saar, Martinez-Gonzalez, Pons-Porteria, Lecture Notes in Physics 665, Springer, 2009

Acknowledgments LV is supported by FP7- PEOPLE-2002IRG4-4-IRG#202182 and MICINN grant AYA2008-03531. I acknowledge the use of the Legacy Archive for Microwave Background Data Analysis (LAMBDA). Support for LAMBDA is provided by the NASA Office of Space Science.

References

1. J. P. Wall, and C. R. Jenkins, *Practical Statistics for Astronomers*, (Cambridge University Press, Cambridge, 2003). 151, 175
2. W. H. Press et al., *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1992). 152, 153, 175
3. M. Kowalski, et al., Astronphys. J. **686**, 749 (2008). 152, 153
4. W. J. Percival, et al., Astronphys. J. **657**, 645 (2007). 152
5. Legacy Archive for Microwave Background Data Analysis http://lambda.gsfc.nasa.gov/ cosmological parameters plotter. 158, 161, 173
6. J. Hamann, S. Hannestad, G. Raffelt and Y. Y. Wong, JCAP **0708**, 021 (2007). 158, 159
7. W. Cash, Astronphys. J. **228**, 939 (1979). 159
8. D. N. Spergel, et al., Astron Phys. J. Suppl. **170**, 377 (2007). 160, 161
9. D. N. Spergel, et al., Astron Phys. J. Suppl. **148**, 175 (2003). 160
10. R.A. Fisher, J. Roy. Stat. Soc. **98**, 39 (1935). 162
11. A. Kosowsky, M. Milosaljevic and R. Jimenez, Phys. Rev. **D66**, 063007 (2002). 163
12. A. Refregier, A. Amara, T. Kitching, and A. Rassat, arXiv:0810.1285 (2008). 164
13. L. Verde, H. Peiris, and R. Jimenez, JCAP **0601**, 019 (2006). 165
14. D. J. Eisenstein, et al. Astronphys. J. **633**, 574 (2005). 166
15. S. Cole et al., MNRAS, 362, 505 (2005). 166
16. W. Percival, et al. Astronphys. J. **657**, 645 (2007). 166
17. H. Seo, and D. J. Eisenstein, Astronphys. J. **598**, 720 (2003). 167
18. H. Seo and D. J. Eisenstein, Astronphys. J. **664**, 679 (2007). 169, 170
19. A. R. Liddle, Mon. Not. Roy. Astron. Soc. **351**, L49–L53 (2004). 172
20. A. O'Hagan, *Kendall's Advanced theory of statistics, Volume 2b, Bayesian Inference*, (Arnold, Huntinston Beach, 1992). 173, 175
21. H. Akaike, IEEE Trans. Auto. Contol. **19**, 716 (1974). 173
22. G. Schwarz, Ann. Stat. **5**, 461 (1978). 173
23. A. F. Heavens, T. D. Kitching, and L. Verde, Mon. Not. Roy. Astron. Soc. **380**: 1029–1035 (2007). 174
24. A. F. Heavens, Lectures given at the "Francesco Lucchin" School of Astrophysics, Bertinoro, Italy, 25–29 (May 2009), arXiv:0906.0664 174
25. L. Verde, XIX Canary Island Winter School "*The Cosmic Microwave Background: From Quantum Fluctuations to the Present Universe*," (Cambridge University press, Cambridge 2009) 174, 176

26. A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics, Volume 1, Distribution Theory*, (Arnold, Huntinston Beach, 1924). 175
27. A. Stuart and J. K. Ord, *Kendall's Advanced theory of statistics, Volume 2a, Classical Inference and Relationship*, (Arnold, Huntinston Beach, 1992). 175
28. V. J. Martinez, E. Saar, O.Lahav, and *Statistics of the Galaxy Distribution*, (Chapman & Hall, New York, 2002). 176

Index

Wolschin, G. (ed.): *Index*. Lect. Notes Phys. **800**, 179–181 (2010)
DOI 10.1007/978-3-642-10598-2