ASPECTS OF MODERN COSMOLOGY

Bruce Bassett

A thesis submitted in partial fulfillment of the degree of Masters of Science.

Department of Mathematics and Applied Mathematics

University of Cape Town

February 1997

ASPECTS OF MODERN COSMOLOGY

Bruce Bassett

Department of Mathematics and Applied Mathematics

University of Cape Town

February 1997

Supervisors:

Professor George F.R. Ellis (Maths & Applied Mathematics)

Professor Tony P. Fairall (Astronomy)

Abstract

The main work of this thesis can be summarised as:

• An implementation of canonical quantisation to the covariant and gauge-invariant approach to cosmological perturbations. Standard results are reproduced. We discuss the advantages of this formalism over non-covariant and non gauge-invariant formalisms.

• A characterisation of linear gravitational waves in a covariant way is achieved. The evolution equations for the electric and magnetic parts of the Weyl tensor are shown to be of different order. In particular, the electric part appears to have a third order evolution equation, while the magnetic part has a second order evolution equation.

• It is shown that the "silent" nature of the evolution equations for for irrotational dust can be extended to the case of vortical dust. This may be relevant for the endpoints of gravitational collapse since the vorticity begins to grow as soon as density contrast becomes non-linear, as is the case in galaxies, showing that the irrotational silent universes are unstable. The main problem in accepting such vortical silent universes lies in proving integrability of the equations which has not been achieved so far, even in the irrotational case.

• A review of issues in the Cosmic Microwave Background (CMB) is given, focussing particularly on points such as ergodicity, decaying modes, foreground contamination, recombination, spectral distortions and polarisation of the CMB.

• A review of methods in gravitational lensing is presented, together with a hierarchy of distance measures in cosmology, forming an introduction to the following two chapters.

• A common belief that photon conservation implies that the all-sky averaged area

distance in inhomogeneous universes must be that of the background, matter-averaged Robertson-Walker area distance is disproven. This means that there will in general be gravitational lensing effects even on large angular scales.

• The realistic situation in which gravitational lensing leads to caustic formation is discussed. It is claimed that this invalidates many accepted beliefs concerning high-redshift observations in inhomogeneous universes. One application of importance is the CMB. Possible implications are discussed.

• Random Gaussian fields are ubiquitous in modern statistical physics, and particularly important in CMB studies. Here we give accurate analytical functions approximating $\int e^{-x^2}\, dx$, the simplest of which is just the kink soliton.

INTRODUCTION

This thesis is a distillation of some of the many interesting projects that I have worked on since my undergraduate days, and presents a hopefully coherent tour of two major themes in modern observational cosmology: the quest for covariant guage-invariant perturbation formalisms and the self-consistent study of the Cosmic Microwave Background (CMB) including the effects of gravitational lensing.

Right from the outset I must apologize to the reader who may be unfamiliar with certain of the topics of the thesis and may not find a satisfactory introduction in the relevant chapters. I have assumed at least a vague familiarity with all the topics covered in the thesis, together with standard big-bang cosmology. To have included thorough introductions to all the chapters would have been verbose, as this has been done much more elegantly elsewhere, and would have turned this work into a threat to the trees everywhere.

The thesis is split into two parts, and is essentially chronological as far as the universe's history is concerned. Cosmology has long struggled to implement the covariance explicit in General Relativity. This is still a problem for cosmology, not only in achieving this goal, but also determining even if this is the right approach to take. However it is a good approach in general and this is the setting for the first part of the thesis. The reader is assumed to be familiar with the basic aspects of the covariant approach, as discussed in, e.g. Ellis (1971) [8]. The thesis starts with a quantisation of the covariant approach to quantum perturbations, which is still a very new field. It then moves on to the related field of covariant characterisation of linear gravitational waves and then to covariant descriptions of the nonlinear epoch of structure formation, silent universes and the question surrounding the importance of vorticity on cosmological scales and the conjecture whether the silent universe formalism can support vorticity.

The next major theme is the cosmic microwave background (CMB). A review of the CMB on all scales is given in chapter 4. This serves as a basis for the following chapters on gravitational lensing in a cosmological context. The main aims are to prove a theorem, regarding the nature of the all-sky averaged area-distance in inhomogeneous universes. Secondly, I explore some of the implications of gravitatonal lensing for the CMB, particularly at small scales when caustics are taken into account. This will hopefully provide a theoretical basis for the great mass of future experimental work to be conducted in the small-scale regions of the CMB sky & which potentially offer the chance to resolve many of the fundmental questions surrounding structure formation.

The final chapter is a digression to numerical analysis inspired by issues in random Gaussian fields. An approximation to $\int e^{-x^2} dx$ is given in terms of the kink soliton and a generalised Maxwell-distribution, which is accurate to less than 0.2%.

All chapters of the thesis are either partially or completely original work, apart from the reviews of the CMB and gravitational lensing in chapters 4 and 5. The work presented in chapters 1, 2, 6 and 7 was done in collaboration with various authors; namely Victor Villalba, Peter Dunsby, George Ellis, Nazeem Mustapha and Charles Hellaby. The work presented in chapter 8 is my own.

Acknowledgements

It is usual for acknowledgements of this sort to start with the recognition of the enormous contribution made by ones parents. My case is no different and it is not possible to express this in a few words, other than to dedicate this thesis to them, though it be a pathetic piece of work compared with theirs.

Similarly I would like to thank my many friends, particularly Josh Bryer, whose company I have enjoyed for over a decade, and Mark Katz.

On the academic side, I want to thank my supervisors Professors George Ellis and Tony Fairall for their help. I particularly apreciate the insight and opportunities which George has provided me from the middle of 1992 onwards, and which continues today. I would also like to sincerely thank the other people with whom I have worked on papers with: Nazeem Mustapha, Charles Hellaby, Peter Dunsby. A special thanks goes to Victor Villalba for his hospitality and his collaboration on the paper from chapter 1, and Pablo Haines for producing the IRAS figures of chapter 1, and the general entertainment he offered. Further, a great deal was learned from a bunch of extremely interesting and intelligent people: the students of the cosmology group at UCT: Tim, Nazeem, Tim, Rodney, Conrad and Nico. Then there are the people who really helped to make my academic life easier, which I greatly appreciate: Di, Jill, Maureen and Arddy Mossop.

Finally I would like to extend a huge vote of thanks to the members of the various sectors of SISSA whom I have had contact with, but particularly the members of the astrophysics sector. Their hospitality and accomodating friendliness was exceptional during a very difficult period when the first draft of this thesis was finished. In particular, I would like to thank Dennis Sciama, John Miller, and Antonio Lanza in this regard. Finally I particularly want to thank Stefano Liberati ("Hey-Hey it's OK") and the members of Rm 105, SISSA in 1996 for their superb friendship. Sincere apologies to everyone else who deserves mention but have been inadvertintly omitted. Please don't hold it against me. After all, this isn't the Oscars.

Conventions

The pure Einstein Field equations are assumed to hold:

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = \kappa T_{\mu\nu}$$

everywhere. Spacetime always has four dimensions. All indices, whether Greek or Latin run from 1 to 4, except in chapter 3 where tetrad indices may only be spatial. The source terms for the field equations are perfect fluids except in special cases where the form of the imperfection is made explicit.

In addition the following acronyms are used often in this thesis:

FLRW	Friedmann-Lemaître-Robertson-Walker spacetimes
GIC	Gauge-Invariant and Covariant
CMB	Cosmic Microwave Background radiation
CDM	Cold Dark Matter
HDM	Hot Dark Matter
MDM	Mixed Dark Matter
PBI	Primordial Baryon Isocurvature
COBE	Cosmic Background Explorer satellite
DMR	Differential Microwave Radiometer
FIRAS	Far Infra Red Absolute Spectrophotometer
(I)SW	(Integrated) Sachs Wolfe effect

Contents

Part I

The Covariant Approach to Cosmology

Chapter 1

Quantisation of Covariant, Gauge-Invariant Cosmological Perturbations

"Come forth into the light of things
Let Nature be your teacher"
W. Wordsworth

1.1 Introduction

The search for quantum gravity is perhaps the holy grail of modern theoretical cosmology, promising to reveal the secrets of the very origin of our universe. A summary of present approaches to this problem, which include string theory, the superspace formulation, Ashtekar and loop variables [5], gravity via random surfaces and so on, is not our aim in this chapter. Rather we wish to discuss a later time in the universe's history, when a plausible assumption is that General Relativity is an accurate low-energy limit of some more fundamental theory of spacetime, in which unification of gravity with the other forces may or may not occur. At times after the Planck era, gravity is split from the other forces even if it were unified at higher temperatures. In particular it is reasonable to assume that spacetime is well described after the Planck time by a manifold structure with a constant topology. This is the epoch of interest in this chapter, the arena where grand unification, baryogenesis and inflation are proposed to take place. The idea that scalar fields were important during this period is one of the fundamental ideas in modern particle physics and early universe cosmology. Inflation and the formation of topological defects via spontaneous symmetry breaking and the Kibble mechanism, offer at present the two most promising mechanisms for generating the primordial fluctuations that lead to the temperature anisotropies recently detected in the cosmic microwave background (CMB) by the *COBE* DMR instrument. On a related but separate tack, the study of initial conditions for the semi-classical regime are determined at the exit of the quantum gravity phase. The Hawking no-boundary proposal [11] and the tunneling boundary condition are two approaches to this problem. Inflation does not usually consider this problem, the usual arguments starting only after the Planckian epoch with the assumption of a region of sufficient homogeneity [1] which can be described by the de Sitter geometry. The classic paper by Halliwell and Hawking [12] attempts to justify the assumption of homogeneity by using the no-boundary proposal to show that the scalar, vector and tensor modes all start in their homogeneous ground states. This has however run into problems since the fundamental requirement that the no- boundary condition actually leads to an inflationary phase, is strongly disputed.

In the end however, semi-classical approaches can only be justified by looking back from a full theory of quantum gravity. Because of the complexity of the problem, studies of the semi-classical regime often focus on different aspects of the problem: either the cosmological side or the quantum field theory side. The first concerns itself with questions of gauge-invariance of perturbations, explicit evolution of the universe and so on, and often uses simple approaches to the quantum field theory involved, neglecting nonlinearity or coupling between fields. The second is more concerned with rigorous analysis of the field theory aspects, including notions of decoherence, noise and the quantum-classical transition. To achieve this the cosmological implications are often neglected, e.g. by using static Minkowski spacetime at zero temperature as background, ignoring the tensor perturbations generated as backreaction on the metric or completely ignoring the gauge problem associated with perturbations in General Relativity. Nevertheless, both approaches are necessary if progress is to be made, so long as the shortcomings of each is recognised. The present work lies in the first category, its focus being the study of a formalism which is both gauge-invariant (GI) and covariant, as opposed to other formalisms which are only gauge- invariant and not covariant or neither.

There are a number of alternatives in the study of quantum modifications of classical behaviour valid under certain assumptions: firstly, minisuperspace models [12] coming from the Wheeler - de Witt equation after freezing out all but two degrees of freedom, and secondly methods based on non-equilibrium field theory, which proceed by defining an order parameter which describes the "coarse grained" evolution of the expectation value of the field. This yields dissipative dynamics from a time-reversible theory. In this case, the infinite degrees of freedom of the field (representing a closed system) are reduced to a few by tracing out the unimportant modes which converts the system into an open one. In this approach [22], the fluctuations in the field are integrated out to get the non-equilibrium effective action.

However, there is another very popular approach to the problem, which can be related formally to the above method: namely that of splitting the field into a homogeneous background (which is treated classically) and an inhomogeneous, anisotropic perturbation, which is then quantised. Unlike in the previous case, it is the perturbations that are of prime importance in this method.

Different formulations have been attempted in this approach [16, 40, 75], but such formulations invariably suffer from one or more drawbacks related to gauge-invariance, physical interpretation or simply the complexity of the calculations that must be performed in realistic cases. In this chapter we consider the quantisation of cosmological perturbations based on the gauge-invariant and covariant (GIC) formalism [19, 35] developed for classical cosmological perturbations.

Put in a crude way, the aims of this work are to establish a quantisation procedure for the covariant approach based on the canonical quantisation picture. We then aim to compare the results with previous methods in the hope that the advantages of the covariant approach will allow discussion of more difficult problems related to multi-fields, nonlinearity and the validity of semi-classical cosmological perturbations in general. As a test case we apply our approach here to a single scalar field.

I would like to thank Victor Villalba with whom the research work in this chapter was undertaken.

1.2 Quantum field theory in expanding and curved spacetimes

In this section we give a brief, mainly qualitative overview of quantum field theory in curved spacetimes. Many excellent reviews are available which make different demands on the background of the reader. Some of the better ones are by Birrel & Davies [30], Brandenberger [2] and Ford [3].

The transition from flat, Minkowski space to a curved background makes the study of second quantisation more interesting. This is essentially due to the equivalence principle and the lack of a privileged observer in General Relativity.

1.2.1 Axiomatic steps to quantisation

There are four steps needed in general to construct a quantum field theory:

- The Lagrangian, Hamiltonian or equation of motion of the classical field.
- A quantisation procedure - canonical or path integral quantisation for example.
- A complete Characterisation of the quantum states.
- Physical interpretation of the states and of observables.

Now in Minkowski spacetime, Lorentz invariance is used at each step. It is used to construct a unique vacuum for example, while in curved or expanding spacetimes we cannot fall back on it as a guide, as mentioned before. However it is possible to pursue formally the first two steps in arbitrary spacetimes without much problem. The real issues of difference arise in considering the last two steps. In a curved spacetime there is no unique vacuum, for example, which physically results in particle production. Another way of stating this is that one cannot make a unique decomposition into positive and negative frequency modes. As a result step four becomes very difficult.

1.2.2 The effective potential

The effective potential is a ubiquitous feature of modern field theory. Here we will concentrate on the physical interpretation of the effective potential rather than on techniques for the summing of loop diagrams to calculate it. For this see [2].

The first major benefit of the effective potential, denoted $V_{\text{eff}}(\phi)$, is that it is the potential energy in quantum field theory - more precisely, $V_{\text{eff}}(\overline{\phi})$ is the minimum of the energy density expectation value in the class of all normalised states $|a>$ satisfying the condition $< a|\phi|a >= \overline{\phi}$. In other words,

$$V_{\text{eff}}(\overline{\phi}) =< a|H|a > \tag{1.1}$$

where H is the Hamiltonian density matrix for the field.

The name "effective potential" is demonstrative of the fact that other fields can contribute to the dynamics and potential energy of the system in question. In particular the temperature (assuming local thermodynamic equilibrium holds) can contribute to the effective potential and indeed this is just the standard view of the way to spontaneous symmetry breaking.

When we discuss the quantisation of a scalar field later in this chapter we will write $V(\phi)$, but are talking of the effective potential of the field.

1.2.3 Finite temperature quantum field theory

Because we are interested in quantisation of a field in the early universe, the assumption that scatterings occur in vacuum, while valid in particle accelerators, is completely invalid here. The standard resolution to this problem is finite temperature field theory which copies thermodynamics in suggesting that scatterings in the early universe take place on a background treated as a thermal bath at temperature T. This thermal bath replaces the vacuum and we replace all our $T = 0$ operators, such as our Green's functions, with their finite temperature equivalents. Note that in gauge theories, the finite-temperature formalism is not gauge-invariant [4].

At finite temperature our operators gain factors such as $\Sigma e^{-\beta H}$ and $\text{Tr}e^{-\beta H}$, where β is the standard inverse temperature parameter, and $\text{Tr}M$ denotes the trace over the matrix M. Now it turns out [2] that the only difference between finite temperature (flat space) field theory and the zero-temperature equivalent is a difference in the boundary conditions on the set of paths on which the measure for the functional integrals has its support. However, this change in boundary conditions is dramatic. Once we Wick rotate to Euclidean space the finite temperature Green's functions become periodic or anti-periodic (for bosonic and fermionic fields respectively) in Euclidean time with period β. This periodicity disappears when we take the zero temperature, $\beta \to \infty$, limit. This implies that only paths in the functional integral which have the same periodicity as the field under consideration, contribute. Further, the periodicity implies that the momentum is discrete in the ik_0 direction, taking values $-i\beta\omega_n = 2\pi n$ for Bose field and $-\beta\omega_n = (2n+1)\pi$ for Fermi fields.

All of this implies a change to the Feynman rules where the loop integrals and vertex δ functions pick up factors of $-i\beta$ and are reduced from four dimensions to three dimensions summed over all (discrete) momentum values, n. The zero and finite temperature Feynman rules are given below ($\delta^{(m)}$ represents the $m-$dimensional Dirac delta).

	Zero Temperature		Finite Temperature
loop integral	$\int \frac{d^4k}{(2\pi)^4}$	$\longrightarrow$	$\frac{1}{-i\beta}\sum_{-\infty}^{\infty}\int \frac{d^3}{(2\pi)^3}$
vertex δ function	$(2\pi)^4\delta^{(4)}(\sum k_i)$	$\longrightarrow$	$\frac{\beta}{i}(2\pi)^3\delta(\sum \omega_i)\delta^{(3)}(\sum k_i)$

Symmetry breaking through finite temperature effects

Consider a scalar field in a thermal bath. Its effective potential (at one loop) is the same as the zero-temperature case but with the addition of an extra term which is a weighted integral over $e^{-\beta E_k}$. What this does in the high-temperature limit is to introduce a $T^2\overline{\phi}^2$ term - which is effectively a (temperature dependent) mass term ! What this implies is that at very high temperature, $T \gg T_c$, where T_c is the critical temperature at which the minima of $V_{\text{eff}}(\overline{\phi})$ become degenerate, $V_{\text{eff}}(\overline{\phi})$ has only a global minimum at $\overline{\phi} = 0$ due to the dominance of the $T^2\overline{\phi}^2$ term. However, once $T < T_c$ the minimum at $\overline{\phi} = 0$ is no longer absolute (the global minimum now being at $\overline{\phi} = \sigma > 0$) but rather represents a metastable false vacuum (see figure 1.3.2).

What is the value of this critical temperature T_c ? It depends of course on which potential one chooses [1] but since the only mass scale in the problem is σ, T_c is generically of the same order of magnitude as σ and $T_c \propto \sigma$.

This symmetry breaking is a vital feature of modern field theory. It appears in the Weinberg-Salam Electroweak theory of $SU(2) \times U(1)$, in many grand unified theories, and in cosmology, where it leads (when causality is imposed) to the production of topological defects via the Kibble mechanism. Further t'Hooft has shown that any theory which exhibits spontaneous symmetry breaking is renormalisable, a very desirable feature. A natural

[1] The "pick a potential - any potential" freedom is one of the weaknesses of inflation.

extension of this brief review would be to consider the decay of the metastable (false) vacuum occurring from symmetry breaking. However, for cosmological interests this leads us into the details of the old and new inflationary paradigms which we want to avoid as it is not relevant to the general issues we want to discuss. Hence we will move straight on to the production of particles in expanding curved spacetimes.

1.2.4 Particle production from the vacuum and Hawking radiation

We have already hinted at the major effect of moving to an expanding or curved spacetime - the effect of particle production. The crucial thing is that spacetime is no longer globally invariant under the Lorentz group because of covariance: there cannot be any *mathematically* preferred frame in GR [2] and this has dramatic consequences. It is this which stops us from splitting uniquely into positive and negative frequency modes, except in asymptotically flat regions or static universes [2]. In this case there exists a timelike Killing vector and the parameter t along this vector field can be used to define the required frequency dependence of the modes in an invariant way.

However in general two observers will do the space-time splitting in a different way. Hence their decomposition of any quantum fields into positive and frequency modes will be different. To quantify this, consider a scalar field $\phi(\mathbf{x},t)$. Observer 1 uses a set of coordinates O_1, defines creation and annihilation operators $a, a^\dagger$ with respect to O_1 and finds:

$$\phi(\mathbf{x},t) = \sum_k [a_k f_k(\mathbf{x},t) + a_k^\dagger f_k^*(\mathbf{x},t)] \tag{1.2}$$

A second observer with coordinate system O_2 defines creation and annihilation operators $b_l, b_l^\dagger$ and decomposes the field as:

$$\phi(\mathbf{x}',t') = \sum_\ell [b_l g_l(\mathbf{x}',t') + b_l^\dagger g_l^*(\mathbf{x}',t')] \tag{1.3}$$

Now in the most general case we must be able to write the positive frequency modes of O_2 as a linear combination of the modes of O_1:

$$g_l = \sum_k (\alpha_{lk} f_k + \beta_{lk} f_k^*) \tag{1.4}$$

with the α_{lk}, β_{lk} known as Bogoliubov coefficients. Now the point of doing this two-observer decomposition is that *if any* of the β_{lk} are non-zero, then there is mixing between the positive and negative frequency modes and observer 2 will see a non-zero number of particles in the vacuum defined by observer 1. The expectation value for this number is:

$$\langle 0|N_k|0\rangle = \sum_l |\beta_{kl}|^2 \tag{1.5}$$

where $|0\rangle$ denotes the vacuum w.r.t. observer 1, and $N_k = b_k^\dagger b_k$ is the number operator for mode k of observer 2.

[2] In cosmology one can use the CMB to define a preferred frame, but it is not a preferred frame of the underlying theory.

In the case of Hawking radiation around a black hole we have [2]:

$$\langle 0|N_k|0\rangle = \left(\exp(\frac{\omega_k}{T_H} - 1)\right)^{-1} \tag{1.6}$$

which is a blackbody spectrum at the Hawking temperature $T_H = H/2\pi$, where H is the Hubble constant.

1.2.5 Production of perturbations in inflation

In the case of an expanding universe, such as the de Sitter background of inflation, where the Hubble constant truly is a constant and the scale factor increases as e^{Ht}, we can use these results on Hawking radiation because of the equivalence principle. The equivalence principle says that effects that occur because of gravitational fields should also occur for accelerating observers. This is seen directly as Unruh radiation for uniformly accelerating observers in the Rindler spacetime for example.

In the de Sitter case, one can follow the steps outlined above and calculate the Bogoliubov coefficients. In addition it can be verified that to first order in H^{-1} the radiation is thermal [2], with an effective temperature $T_H = H/2\pi$. It is this thermal flux of "particles" that is the way inflation produces a spectrum of energy density perturbations. Since H is constant in the pure de Sitter phase, the amplitude of perturbations, which is fixed as their scale crosses that of the Hubble radius (not the horizon as is often said), is (almost) independent of wavelength in this case because of the almost constancy of the Hubble constant in inflationary models. In classical de Sitter spacetime the Hubble constant is truly constant. However, the draining backreaction of the Hawking radiation at one loop level causes a variation in H even in this case which lightly breaks the scale-free nature of the spectrum. However, it is this fact that gives rise to the famous Harrison-Zel'dovich scale-invariant spectrum of perturbations:

$$P_\phi(k) = Ak^n \quad , \quad n \simeq 1 \tag{1.7}$$

This has been validated to some extent by analysis of the COBE data which finds $n \simeq 1.1 \pm .3$ is the most favoured value, once averaged over all the different approaches to obtaining the spectral index n.

One point to make is that Hawking radiation comes from vacuum fluctuations in a non-trivial background geometry *with a horizon*. There is another effect which may also have been very important: production of particles by polarisation of the vacuum by strong and rapidly changing gravitational fields. This is analogous to squeezed-state production of particles in an external electromagnetic field. This was investigated thoroughly after Misner's chaotic cosmology program was introduced, the idea being that anisotropy of the gravitational field was the source of particle production which thereby reduced the anisotropy of spacetime, an effect it was hoped could explain the observed isotropy of the universe. This effect can be thought of as introducing a non-zero vacuum expectation value for the stress-energy tensor, an effect that has been modeled extensively in classical cosmology by adding a bulk-viscous term to the stress-energy tensor.

This method of producing particles has been thoroughly studied by Grischuk in recent years [40], who has shown that a significant spectrum of perturbations can be achieved if there is a sudden transition from a de Sitter to FLRW phase. However, it should be emphasised that at present the distinctions between Hawking radiation and Squeezed state production are not clear, with the strong possibility that they are different facets of the same phenomenon. If they are different, then the biggest problem for the squeezed state process is to produce a nearly scale-free spectrum over a large number of decades in scale.

1.3 Overview of the gauge, background splitting and covariance problems

We now move on to motivations for the original work to be presented in this chapter.

1.3.1 The gauge problem

The gauge problem in cosmology is well known, but fairly subtle. It can be viewed either from a passive or an active point of view. In the active view [19] the gauge is viewed as a mapping (identification) of worldlines and spatial sections from the background into the real, inhomogeneous geometry. By changing the mapping between the two geometries, the perturbation changes (see figure 1). For example, if we identify surfaces of constant density in the background with surfaces of constant density in the real spacetime, then the energy density perturbation: $\delta\mu = \mu - \overline{\mu} = 0$ vanishes. Thus by changing the gauge, an unphysical operation, we may make the density perturbation appear to disappear. The passive view of the gauge-problem is to view it as a coordinate problem: by changing the coordinates one is using infinitesimally, the values of tensor fields on spacetime change. The standard lore is that the gauge issue is only important on superhorizon scales, since on these scales it is not possible to correlate observational quantities via causal processes. Conversely, on sub-horizon scales the gauge problem is usually ignored. In practise this is correct since the coordinates used are smooth. However, in principle one could make a large (i.e. not infinitesimal) change to coordinates which vary extremely rapidly so that observational quantities would depend crucially on the gauge even on sub-horizon scales.

By considering general infinitesimal coordinate transformations one may construct linear combinations of tensors which are invariant under the group of such transformations. This is the approach pioneered by Bardeen (1980) [15] and summarised and extended in [16]. In practice, most calculations are done in a specific gauge (such as the longitudinal gauge) to simplify the calculations, thus losing covariance. In addition, the Bardeen variables are only gauge-invariant at first order. The only alternative prior to this was to use gauge-dependent variables but to specify all the gauge-freedom (and hence eliminate gauge-modes) by choosing a particular gauge. This has the advantage that one can allow the "physics to choose the gauge" and simplifies the calculations, but makes the comparison between work in different gauges extremely difficult. The other approach is to use the lemma (see e.g.

[18]) that any tensorial object which vanishes in the set of background geometries under consideration (usually the FLRW geometries), will be a GI variable in the real spacetime. This approach can be used on any space with some symmetry and is not explicitly linear. Examples of GI variables using the FLRW models as backgrounds are: the shear σ_{ab}, the vorticity, ω_{ab}, the electric, E_{ab}, and magnetic H_{ab}, parts of the Weyl tensor [34] and the set of covariant GI perturbation variables developed since 1989 that are the basic variables for this chapter [19] and the accompanying paper [6].

1.3.2 The background splitting problem

The usual approach to semi-classical perturbations is to consider a *gauge-dependent* perturbation of the field, $\phi(\mathbf{x},t)$:

$$\phi(\mathbf{x},t) = \phi_0(t) + \delta\phi(\mathbf{x},t) \tag{1.8}$$

Apart from being gauge-dependent, this method of splitting quantum fields into a homogeneous background and a small perturbation, which is then quantised, is fraught with technical problems, as discussed in detail by Guth and Pi (1985) [21] and [22]. Since studies of inflationary perturbations have mainly been studies in this framework, these problems are naturally inherited by inflation. Here we give a brief review of these problems:

(1) Inflation relies on the slow roll picture in which the scalar field "rolls" down a gentle gradient potential, thus allowing sufficient inflation. However, at early times when the temperatures are very high, the statement $\Phi \simeq 0$ cannot be true. This is because the fluctuations in the field will be very large. The only constraint that one can impose is that the spatial or time averages of the field are nearly zero. As cooling occurs it is likely that the large fluctuations will cause the field locally to find the global minima very quickly (see figure 2), thus avoiding the slow roll regime and violating the crucial constraint that $(\nabla\phi)^2 \ll V(\phi)$, even though suitable averages of the field might evolve slowly. In fact, this is the basis of Linde's chaotic universe idea [1]. He realised that although the mean field may drop below the threshold required for inflation, there will always be spatial regions where the quantum fluctuations are large enough to begin inflation of that region again. This stochastic process then continues for ever, leading to Linde's proposed fractal structure for the universe.

(2) In pure de Sitter space, how does the background field know how to choose a four velocity, since de Sitter space has no preferred frame of reference since it is maximally symmetric. Further, static de Sitter space has an Euler-characteristic of 2 while FLRW solutions have an Euler-characteristic of 0. Since this is a topological invariant, it is not obvious that there is a smooth evolution from one to the other at reheating. The first problem can be overcome by including perturbations since this allows one to define a four velocity as is done in this chapter, proportional to the gradient field of the scalar field ϕ.

(3) It has been shown that the function $\phi_0(t)$ begins in a thermal ensemble with an exact symmetry $\phi = -\phi$ [21]. Thus the expectation value of the field, $\langle\phi\rangle$, is always zero. Thus

the physical meaning of the background ϕ_0 is obscure.

(4) The standard finite-temperature field theory is not sufficient to describe the slow roll in inflation shown in fig.(1.3.2). Full non-equilibrium quantum field theory is required to describe the mean behaviour of the field [22] in many cases. The approach of background splitting was found to be reasonable by Guth and Pi [21] in practice, and can be translated as the fact that the quantum fluctuations of the full quantised field is roughly the mean behaviour of the perturbations when treated as a field of their own and quantised. However, this picture can fail dramatically in some cases [22]. This is because dissipation is non-Markovian and because the standard effective-potential formalism assumes at least local equilibrium, which is needed to define a temperature. Since this is not true in the phase transitions fundamental to the inflationary picture and modern particle physics, the results gathered using it must be treated with care.

(5) Another highly non-trivial issue is the quantum-to-classical transition of the inflationary perturbations. One method to affect this, due to Hu *et al* [23], is to split the field into a homogeneous component (the order parameter) and a high-frequency noise. The order-parameter is defined as a suitable volume integral of the field, while the noise acts to convert the closed system into an open one, causing decoherence and classical energy density perturbations to appear. While this is well-defined in studies of condensed matter systems, the plausibility of marrying this method with the requirements of gauge-invariance is far from obvious and as yet untouched. Since the order-parameter is defined as a field average we are faced with the same dilemma as occurs in classical cosmology: the need for a covariant and gauge-invariant averaging procedure.

1.3.3 The covariance problem

The covariance problem is simply that ideally we would like to do all our calculations in an explicitly covariant way. Specifying gauge-freedoms, or even using the GI Bardeen approach does not allow us to do this. The use of potentials, as in the Bardeen approach in particular, is nonlocal in the sense that changing boundary conditions outside our past null cone leads to changes to the GI metric potentials inside the past null cone.

Further, the ability to use any coordinate system to calculate details of the theory is highly desirable.

Finally, there may be a deeper reason for requiring covariance. Since Hawking radiation and the lack of a physically invariant vacuum is essentially a result of the covariance of full General Relativity, a covariant perturbation formalism should, at the very least, give rise to clarifications of results obtained in non-covariant approaches.

1.4 The covariant approach to cosmology

The covariant approach relies on a choice of observer fluid four-velocity, u_a. By splitting the covariant derivative as [8]:

$$u_{a;b} = \sigma_{ab} + \omega_{ab} - \dot{u}_a u_b + \frac{1}{3}\Theta h_{ab} \tag{1.9}$$

where $h_{ab} = g_{ab} + u_a u_b$ is the projection tensor and hence the metric in the spacelike hypersurfaces orthogonal to u_a ($\omega_a = 0$). In this equation, σ_{ab} represents the volume preserving shear, ω_{ab} the vorticity and $\Theta \equiv u^a_{;a}$ is the expansion of the flow. One may define a scale factor, S, (which coincides with the FLRW one in that case) via:

$$\frac{\dot{S}}{S} = \frac{1}{3}\Theta \tag{1.10}$$

Using the conservation equations $T^{ab}{}_{;a} = 0$ projected parallel and orthogonal to the fluid flow, one obtains the fully nonlinear, covariant fluid evolution equations (see Ellis, 1971 [8]). These form the basis of the nonlinear "silent" universe approximation and do not specify any background spacetime. The method then proceeds by finding simple, physical, variables which are gauge-invariant (GI) to all orders. Exact, covariant evolution equations for them are then found from the fluid equations. These equations are completely non-linear and are ***subsequently linearised about given background models***, such as the FLRW or Bianchi universes. We will contrast the covariant approach with the Bardeen and other non-GI formalisms often in this chapter. For a thorough discussion of the covariant approach we refer the reader to earlier work aimed at establishing the method [19, 36, 36, 39] and its relation to the Bardeen approach and the Newtonian analogue [35, 20].

Following the lemma on gauge invariance mentioned before, we see that in FLRW backgrounds, the following dynamical variables are explicitly GIC since they vanish for FLRW models:

$$\sigma_{ab} = \omega_{ab} = {}^{(3)}\nabla f = 0 \tag{1.11}$$

where f is any covariantly defined scalar field on the FLRW background, such as the energy density or pressure, and we use the shorthand ${}^{(3)}\nabla^a$ to denote the full covariant derivative projected into the spatial slices via contraction with h_{ab}.

The archetypal scalar density GIC variable in this approach is Δ, defined by the hierarchy:

$$\Delta \equiv S^{(3)}\nabla^a \mathcal{D}_a \ , \quad \mathcal{D}_a \equiv \frac{S^{(3)}}{\mu} \nabla_a \mu \tag{1.12}$$

where μ is the energy density, S is the scale factor of the universe and ${}^{(3)}\nabla_a \equiv h^c_a \nabla_c$ is the covariant derivative projected into the spatial hypersurfaces. $\mathcal{D}_a$ is dimensionless and represents the fractional, comoving spatial density gradient of the fluid. Note that both $\mathcal{D}_a$ and Δ are GIC in FLRW or Bianchi backgrounds because they vanish there. This is true for any type of matter, and hence the quantisation procedure presented here can be extended unchanged to different cases of interest. There are two possible choices for the scalar GIC variable describing density fluctuations, Δ and $\mathcal{D} \equiv (\mathcal{D}_a\mathcal{D}^a)^{1/2}$. The second is a fairly obvious choice, since it measures the magnitude of the comoving spatial density gradient, but why use Δ ? It has been shown [13] that to first order

$$\Delta_k \simeq -k^2\epsilon_m^{(k)} \tag{1.13}$$

where the label k denotes the Fourier harmonic component and ϵ_m is the GI (at first order) Bardeen variable which in the comoving gauge or on small scales, coincides with the usual density contrast. More intuitively, Δ gives information about fluid dynamics. Since it is a divergence, the integral of Δ over a sphere of radius R will give the spatial flux of matter into or out of the sphere, and tells one about the clustering or the formation of voids, depending on the sign of Δ. More particularly, it is closely related to $\sigma_R^2 = \langle(\delta M/M)^2\rangle$, the mass variance in spheres of radius R, once ensemble averaging has been done.

1.4.1 Comparison of Δ, $\mathcal{D}$ and δ for IRAS galaxy data

In this section we would like to use the density field derived using the POTENT [3] analysis of the IRAS [4] galaxy survey to compare the usual density contrast variable, $\delta = \delta\rho/\rho$, which is not gauge-invariant, with the corresponding density variables in the GIC approach.

Although the covariant variables offer significant advantages over traditional ones for analytical treatments of e.g. density waves and segregation of multiple fluids, it is not obvious what they look like for real data. In figures (1.4,1.5) and (1.6) we plot Δ, $\mathcal{D}$ and $\mathcal{D}_a$ projected onto the supergalactic plane, together with the conventional density contrast δ in figure 1.4.1. Since they have to be calculated using derivatives, which are very noisy numerically, the plots of the covariant variables are not as smooth as that for δ. However we can see that Δ behaves as expected: $\Delta \propto -\delta$.

Geometric GIC variables

Δ is a useful GIC *kinematic* variable. A GIC *geometric* variable is C, giving the deviation of the surfaces of constant time from constant curvature manifolds:

$$C = S^{(3)}\nabla^a C_a\,, \quad C_a = S^{(3)}\nabla_a{}^{(3)}R \tag{1.14}$$

where for a *classical* scalar field, ${}^{(3)}R = 2[-\frac{1}{3}\Theta^2 + \sigma^2 + \frac{1}{2}\dot{\phi}^2 + V(\phi)]$ is the Ricci curvature scalar of the three-spaces orthogonal to u^a. It turns out that on superhorizon scales in flat models, C is conserved, so that it can be used to match the spectrum of perturbations as is

[3]The POTENT method is based on the fact that *if the velocity field is irrotational,* it can be derived completely from the gravitational potential. Conversely, if one has the peculiar velocity field then one can derive the underlying density field. The peculiar velocity field can be derived using e.g. the Faber-Jackson or $D_n - \sigma$ relations.

[4]The IRAS data is from a spectroscopic infra-red survey of more than 2500 galaxies of reasonable flux ($> 1.9Jy$) at a wavelength of $60\mu m$. The survey covers more than 88% of the sky and consists almost exclusively of spiral galaxies. The use of the infra-red band means that our galaxy presents much less of a problem than it is in the visible band, where there is a lot of obscuration.

done in standard inflation. For more details on the physical interpretation of the covariant variables see Bruni *et al* [35].

1.4.2 Scalar field and inflation

As we discussed earlier, scalar fields are vital to modern cosmo-particle physics. They are also simple to discuss. As a rich illustrative example, we therefore consider a scalar field minimally coupled to gravity in an expanding FLRW background. We do not treat the conformally coupled or higher order gravity cases here but note that they can be treated in a natural way using an imperfect fluid formalism [5] instead of the present perfect fluid approach. The Lagrangian density for the problem is:

$$\mathcal{L} = -\sqrt{-g}[\frac{1}{2}\nabla_a\phi\nabla^a\phi + V(\phi)] \tag{1.15}$$

with $V(\phi)$ the effective potential describing the mass and self-interaction of the field. ∇^a is the full (unprojected) covariant derivative w.r.t. the metric $g_{\mu\nu}$. The energy density given by $\mu = \frac{1}{2}\dot{\phi}^2 + V(\phi) + \frac{1}{2}(\nabla\phi)^2$ By choosing the four velocity proportional to the normals to the surfaces $\{\phi = const.\}$, $u^a = -(\dot{\phi})^{-1}\nabla^a\phi$, the scalar field takes the form of a perfect fluid, represented by an equation of state $p = (\gamma - 1)\mu$. Following [13] we define the momentum field:

$$\psi \equiv \dot{\phi} \tag{1.16}$$

and from this the dimensionless GIC variable

$$\Phi_a \equiv \frac{S}{\dot{\phi}}{}^{(3)}\nabla_a\psi \tag{1.17}$$

which is the comoving dimensionless spatial gradient in the field ϕ. Then in this case the fundamental density gradient is:

$$\mathcal{D}_a = \frac{\psi^2}{\mu}\Phi_a = \gamma\Phi_a \tag{1.18}$$

The linearised equation of motion for Δ in the case of a classical scalar field has been previously derived [13], and is:

$$\ddot{\Delta} + \mathcal{A}\dot{\Delta} - \mathcal{B}\Delta -{}^{(3)}\nabla^2\Delta = 0 \tag{1.19}$$

where

$$\mathcal{A} \equiv (\frac{5}{3} - \gamma)\Theta - \frac{\dot{\gamma}}{\gamma} \tag{1.20}$$

$$\mathcal{B} \equiv (1 - \frac{\gamma}{2})[\Theta^2(\gamma - \frac{2}{3}) + 9\gamma\frac{K}{S^2}] + \Theta\frac{\dot{\gamma}}{\gamma} \tag{1.21}$$

where $K = \pm1, 0$ is the FLRW curvature constant, and

$$\frac{\dot{\gamma}}{\gamma} = \Theta(\gamma - 2) - \frac{2}{\psi}\frac{\partial V}{\partial\phi} \tag{1.22}$$

[5] i.e. by including fluxes and anisotropic stresses.

Here prime denotes derivative with respect to ϕ. Next it will be useful to change to conformal time, η. This eliminates the S^{-2} dependence from the ${}^{(3)}\nabla$ term after Fourier expansion, yielding:

$$\Delta'' + (\mathcal{A}S - \frac{S'}{S})\Delta' - S^2\mathcal{B}\Delta + k^2 = 0 \tag{1.23}$$

Finally, we will find it necessary to make a change of variable to:

$$\Delta(\mathbf{x}, t) = X(\mathbf{x}, t) \exp(-\frac{1}{2}\int \mathcal{A}(\xi) d\xi) \tag{1.24}$$

which eliminates the Δ' term from the evolution equation, which is then converted to:

$$X'' - m^2(\eta)X = 0 \tag{1.25}$$

with

$$m^2(\eta) = k^2 - S^2\mathcal{B} - \frac{(S^2\mathcal{A} - S')^2}{4S^2} - \frac{1}{2S}(S' - S^2\mathcal{A})(\frac{S'}{S} - S\mathcal{A})' \tag{1.26}$$

where $'$ denotes derivative with respect to conformal time $d\eta = dt/S(t)$. The reward for our pains is that eq. (1.25) is just the mode-expanded Klein-Gordon equation *in Minkowski space* with a time-dependent mass, $m^2(\eta)$, occurring with the opposite sign to the usual Klein- Gordon mass term. Further it is qualitatively identical to results previously obtained (see e.g. [16]) and allows quantisation to proceed via the standard free-field formalism.

1.5 The quantisation procedure

The traditional approach [16] to studying this problem is to write down the Hilbert action for gravity and the action for the source fields of matter, S_m, such as scalar fields or ordinary hydrodynamical matter:

$$S = \int R\sqrt{-g}\; d^4x + S_m \tag{1.27}$$

where $g = det(g_{\mu\nu})$. The metric and scalar field, ϕ, for example are then perturbed

$$g_{\mu\nu} = \eta_{\mu\nu} + \delta g_{\mu\nu}\;, \qquad \phi(\mathbf{x}, t) = \phi_0(t) + \delta\phi(\mathbf{x}, t) \tag{1.28}$$

The problem is that these perturbations are not gauge-invariant. However, it is possible to expand the total action up to second order and write the results in terms of GI linear combinations of separately gauge-dependent quantities. The action is then varied and the classical equation of motion for the GI linear combination derived [16]. This is then postulated to be an operator and the quantisation follows as described below.

In our case, there is no lagrangian or action to start with, simply because there are no explicit metric perturbations. The classical equations of motion for the *gauge invariant* variables are derived directly from the field equations in the covariant approach. However, since we are considering perturbations of a minimally coupled scalar field, there is a relation between Δ and the perturbed field, $\delta\phi(\mathbf{x}, t)$. It is partly described by

$${}^{(3)}\nabla^a\mu = 2\dot{\phi}_0 {}^{(3)}\nabla^a(\delta\dot{\phi}) + V'(\phi){}^{(3)}\nabla^a(\delta\phi) \tag{1.29}$$

where here V' denotes $dV/d\phi$. This equation relates the GIC energy density gradient field explicitly to gradients in the scalar field perturbation and its time derivative. Thus the quantisation of $\delta\phi$, as usually explicitly considered, is implicitly involved during the quantisation of Δ which we consider.

For brevity, in treating the quantum problem, we consider scalar perturbations related to energy density fluctuations. Formally, we make both Δ, corresponding to the GIC fluctuations in the scalar field energy density, and X, operators: $\hat{\Delta}$ and $\hat{X}$. We now proceed with the standard quantisation of X via expansion in eigenfunctions, Q_k, of the Laplace-Beltrami operator ∇^2, on the background spatial geometry. In the case of a flat background, $k = 0$, (not to be confused with the eigenvalue label k, used hereafter) as usually considered in quantum perturbation discussions, $Q_k(\mathbf{x}) = (2\pi)^{-3/2} exp(i\mathbf{k}\cdot\mathbf{x})$ - a basis of plane waves. In the case of open geometries, the eigenfunctions are complicated hyperbolic functions [17]. The operator expansion, over creation and annihilation operators, $a_k^\dagger, a_k$, is then: [30, 16]

$$\hat{X} = \frac{1}{\sqrt{2}}\int dJ[u^*(\eta)Q_k(\mathbf{x})a_k + u(\eta)Q_k^*(\mathbf{x})a_k^\dagger] \tag{1.30}$$

where $dJ = d^3k$ when $K = 0$. We impose the standard commutation relations for bosons:

$$[a_k, a_{k'}] = [a_k^\dagger, a_{k'}^\dagger] = 0,\ [a_k, a_{k'}^\dagger] = \delta_{kk'} \tag{1.31}$$

The normalisation condition for eq.(1.30) then gives [30, 31]

$$\dot{u}_k(\eta_0)u_k^*(\eta_0) - \dot{u}_k^*(\eta_0)u_k(\eta_0) = 2i \tag{1.32}$$

where η_0 is the initial conformal time. The next step is defining positive- and negative-frequency modes and hence a vacuum state $|0\rangle$ with $a_k|0\rangle = 0\ \ \forall\ k$. It is here we want $\hat{X}$ to operate: on the Hilbert space, $\mathcal{H}$, of Fock states. To discuss the time evolution of the amplitude $u_k(\eta)$, as required to analyse the amplitude, rate and spectrum of particles produced from the vacuum, we need to define the vacuum corresponding to the initial and final states properly [30]. This has been done in exact de Sitter spacetime [31] and can in theory be done in perturbed RW models, although technically difficult. Here we limit ourselves to a discussion of the power spectrum and leave this aspect of the problem to future work.

In the spatially flat case, we derive the equation of motion for the amplitudes by substituting this expansion over eigenfunctions into eq.(1.25) and using the result that the projected Laplacian, ${}^{(3)}\nabla^2$ becomes $-k^2/S^2$. It is given by:

$$\ddot{u}_k + (\frac{k^2}{S^2} - m^2(\eta))u_k = 0 \tag{1.33}$$

To make the discussion above about the vacuum more concrete, we consider the initial conditions for the above equation. The normalisation condition is satisfied by any vacuum with $u_k(\eta_0) = c_0k^{-\alpha}$, $\dot{u}_k(\eta_0) = ic_0^{-1}k^{\alpha}$, constant α, c_0. However for de Sitter spacetime the appropriate initial conditions following from eq.s(1.32,1.33) so as to recover the Minkowski limit at small scales are $u_k(\eta_0) = \frac{c_0}{\sqrt{k}} = u_k^*(\eta_0)$, $\dot{u}_k(\eta_0) = \frac{i\sqrt{k}}{c_0}$.

1.6 Comparison with observations - the power spectrum, and the Quantum to Classical Transition

In order to make contact with observations of e.g. the cosmic microwave background (CMB) or galaxy statistics we require the power spectrum of the energy density fluctuations implied by this work, specialised to a flat geometry. However since we are considering quantum perturbations in this work, where we have a superposition of possible states, we must first overcome the problem of the decoherence of the quantum states during the quantum to classical transition. This process is the one in which physical inhomogeneities are created and lead to the temperature anisotropies in the CMB. This is a non-trivial problem [23] and is where, the standard formalism has an advantage. In the standard (non - GI) splitting of the field into two parts, one transforms the system from a closed one, into an open one, where the high-frequency space-dependent component $\delta\phi$ acts as a noise term causing decoherence of the mean field, leading to the classical transition.

In the covariant approach we are somewhat removed from the field itself. Ideally we would like to understand what rôle spatial gradients in the field play in causing decoherence. This is intimately related to the interpretation of our results here. We have elevated our gradient variables to operators, but in what sense are we considering them as small quantum fluctuations about a background ? Ours is a more abstract formulation, and hence an important way of understanding the nature of this quantisation is to compute the power spectrum of quantum fluctuations.

Alternatively one can use a density matrix formulation of the problem [42]. In this framework, the requirement for a quantum system to behave as a classical one is that the trajectory converge to a classically allowed one, *and* that all off-diagonal elements of the density matrix tend to zero, implying decoherence. However, a formulation of the covariant approach to perturbations in terms of a density matrix has not been done yet.

Leaving this aside, the power spectrum is simply the Fourier transform of the two-point correlation function, or rather, the expectation value of this function. If the fluctuations are Gaussian, then these provide a complete description of the statistics of the field [6]

The averaged two-point correlation function is defined by the expectation value [16]:

$$\xi_{\hat{\Delta}}(\eta, \mathbf{r}) = \langle 0|\hat{\Delta}(\eta, \mathbf{x}+\mathbf{r})\hat{\Delta}(\eta, \mathbf{x})|0\rangle \tag{1.34}$$

To be precise this should be defined in terms of the auxiliary variable X, however it differs from Δ only by a time-dependent factor and so does not alter spatial statistics such as the power spectrum, but only their time-evolution, which we are not particularly interested in.

If ψ is homogeneous then this implies that the correlation function is independent of $\mathbf{x}$. This is an assumption in classical theory usually ascribed to statistical isotropy and homogeneity of the perturbations. The power spectrum, $|\delta_k|^2$, is then related to the expectation

[6] Topological defect models do not yield Gaussian fluctuations since the phases are correlated. Hence higher order correlation functions, or fractal statistics, are required.

value of the two-point correlation function via:

$$\xi_{\hat{\Delta}} = \int_0^\infty \frac{dk}{k}\frac{sin(kr)}{kr}|\Delta_k|^2 \tag{1.35}$$

$|\Delta_k|^2$ measures the squared amplitude of fluctuations on a *comoving* scale k. If the fluctuations are Gaussian, the modulus is sufficient since the phases are uncorrelated. By substituting eq.(1.30) into this equation and using the properties of the ladder operators we obtain [16]:

$$|\Delta_k(\eta)|^2 = \frac{1}{2\pi^2}k^3|u_k(\eta)|^2 \tag{1.36}$$

the power spectrum being a function of the conformal time, η. The standard result in inflationary perturbations is that the power spectrum of metric fluctuations on superhorizon scales is scale invariant, being simply proportional to the square of the Hawking temperature $H/2\pi$ [16]. However, $u_k(\eta)$ has a $1/\sqrt{k}$ dependence in the limit $k \to 0$ which means that the spectrum at large scales varies as

$$|\Delta_k|^2 \propto k^2\ , \quad k \ll 1 \tag{1.37}$$

This is still a power-law, but as expected (since it involves gradients of the energy density) it is no longer scale-invariant.

1.6.1 Spectrum of Entropy perturbations

Since the equation of state for a scalar field varies with spatial position, entropy perturbations exist. We define a GIC entropy variable $\mathcal{E}_a$ via [35]:

$$\mathcal{E}_a = \frac{S}{p}\left(\frac{\partial p}{\partial s}\right)^{(3)}_{(\mu)} \nabla_a s \tag{1.38}$$

where s is the entropy density. We then define a GIC entropy scalar perturbation as usual via:

$$\mathcal{E} = S^{(3)}\nabla_a \mathcal{E}^a \tag{1.39}$$

Then in the case of a minimally coupled scalar field we have: [35]

$$\mathcal{E} = (1 - c_s^2)\mu\Delta \tag{1.40}$$

where $c_s^2 = (\partial p/\partial \mu)_s = \dot{p}/\dot{\mu} = \gamma - 1 - \dot{\gamma}/(\Theta\gamma)$.

Now the power spectrum of entropy perturbations can be derived from that of Δ using the fact that the Fourier transform of a product of functions is the convolution of the individual Fourier transforms. Thus:

$$\mathcal{E}_k = \mu(1 - c_s^2)_k * \Delta_k \tag{1.41}$$

and the power spectrum is:

$$|\mathcal{E}_k|^2 = |\mu(1 - c_s^2)_k * \Delta_k|^2 \tag{1.42}$$

where $*$ denotes convolution. Thus we see that the power spectrum of entropy perturbations is related to the power spectra of density perturbations and fluctuations in the speed of sound, c_s^2, in a highly non-trivial way.

1.7 General Considerations

We pointed out earlier that the linearised equations for Δ correspond indirectly to the linearly perturbed part of the field, $\delta\phi$. However, the fully nonlinear equations can be used to describe *any* field interacting with gravity (including backreactions on the gravitational field) as long as there exists a fluid description for the field. Thus there exists a great deal of scope in using the covariant approach to study nonlinear field dynamics, as well as conventional nonlinear density evolution appropriate at late times.

In general the full nonlinear equations of motion in the classical covariant approach for a stress tensor with shear, vorticity and anisotropic stress will involve a large set of GIC variables. The result is that either in the perfect fluid nonlinear case or in the more general linear case, the evolution equations will be coupled.

1.8 Invariance of the method

We have discussed the quantisation of a scalar field, mainly because of its importance in discussions of the early universe and inflation. However, the formalism developed here is equally valid for any other interesting field. This is mainly due to the fact that the canonical GIC variables of the Ellis-Bruni formalism all have the form of a spatial gradient (modulo inessential factors). This yields governing equations of motion which are all second order linear ordinary differential equations. When it is possible to do this, the quantisation procedure then follows through as discussed earlier in this chapter.

1.9 Comparison with other formalisms

The work by Mukhanov [16] identified a single GI scalar variable, which was a linear combination of metric and scalar field perturbations. Here we too have a single scalar variable, but this time there is no difficulty in identifying the physical meaning.

The fact that the method is fully nonlinear at the classical level implies that there may be some way to extend the formalism further back in time, when the fluctuations may have been non-linear. This will be discussed in depth in a later section.

The method offers many advantages over the standard formalism: (1) It is both covariant and gauge-invariant, (2) identification of the various possible GIC variables is trivial, no searching for GI combinations of non-GI variables is needed, (3) metric perturbations are not required, simplifying the work, (4) the *fully non-linear* equations are available for the operators, (5) the quantisation method applies almost unchanged (except e.g. for the details of the equation governing the evolution of u_k) to other perturbations of interest, (6) there is no nonlocal splitting into scalar, vector and tensor as in the Bardeen formalism. This ensures uniqueness of the present decomposition [41].

1.10 Parametric Reheating at the end of inflation

Here we begin our study of nonlinear, non-perturbative and topological aspects of semi-classical quantum gravity.

The traditional view of inflation was that quantum fluctuations in the inflaton field ϕ gave rise to classical perturbations essentially through Hawking radiation - particle production because of the lack of a unique vacuum in curved spacetime. At the end of inflation when reheating occurs via a second order phase transition [7] the inflaton particles are a classical coherent scalar field and the field had reached the minimum of its potential, it executed periodic oscillations around the bottom of the potential and assuming that it coupled non-gravitationally to other fermionic (through interaction terms such as $-h\overline{\psi}\psi\phi$) and/or bosonic fields (with interaction terms such as $-\frac{1}{2}g^2\phi^2\xi^2$), decayed into these particles by scattering, establishing equilibrium. In this way the universe was believed to have been reheated and the huge entropy of the universe created.

However, it was noticed by Kofman *et al* [26] that there can exist a very powerful precursor to this thermalisation - that of non-perturbative parametric production of particles, a process far out of equilibrium. Consider the simplest model of this phase of the inflationary model: the inflaton, treated now as a classical field, is coupled to a light scalar field ξ with coupling constant g. The simplest effective potential of the inflationary field to demonstrate this effect is $V(\phi) = m_\phi^2\phi^2/2$. Then the equation for the evolution of the modes of the field ξ is given by: [26]

$$\ddot{\xi}_k + \Theta\dot{\xi}_k + \left(\frac{k^2}{S^2} + g^2\Phi^2 sin(m_\phi t)\right)\xi_k = 0 \tag{1.43}$$

where the sinusoidal term is due to the coupling to the inflaton field which is executing oscillations at the bottom of the effective potential with amplitude Φ, a slowly decaying function of time. Now if we follow [26] and neglect, as a first approximation, the expansion of the universe, treating S as a constant, $(\rightarrow \Theta = 0)$ then we can recast this equation into that of Mathieu's equation:

$$\xi_k'' + [\alpha(k) + \beta cos(2z)]\xi_k = 0 \tag{1.44}$$

where the new independent variable is $z = m_\phi t$ and $\alpha(k) = k^2/m_\phi^2 S^2 - \beta$, and $\beta = -g^2\Phi^2/8m_\phi^2$. Mathieu's equation has long been known in classical mechanics and is a special case of Hill's equation in the study of Floquet theory - the study of ODE's with periodic coefficients. The vital property of Mathieu's equation is that it possesses bands of exponentially unstable solutions: $\xi_k \propto exp(\mu_k^{(n)} z)$ depending on the parameters $\alpha(k)$ and β. Thus for certain modes, coupling constants and amplitudes of oscillation of the ϕ-field, the number of ξ particles will grow exponentially: $n_k(t) \propto \xi_k^2$. This catastrophic particle production would, for favourable parameter values, strip the inflaton field of its energy within a

[7] A second order phase transition is one in which ϕ changes its value continuously. In a first order phase transition by contrast, ϕ jumps discontinuously at the transition. A first order transition is typically one in which bubbles of new vacuum form and grow within the old vacuum.

few oscillations of ϕ, until the decay rate rate falls below the expansion of the universe (small β) or the backreaction of the large quantum fluctuations shuts off the resonance (large β).

This very explosive production of particles has been called *preheating* by [26]. Note that after preheating, standard reheating and thermalisation is thought to occur via processes $\xi \rightarrow \kappa\kappa$ and $\xi \rightarrow \psi\psi$ where κ and ψ are further bosonic and fermionic fields respectively. The decay rates for these processes being:

$$\Gamma(\xi \rightarrow \kappa\kappa) = \frac{\tilde{g}^4\sigma^2}{8\pi m_\xi}\ , \quad \Gamma(\xi \rightarrow \psi\psi) = \frac{h^2 m_\xi}{8\pi} \tag{1.45}$$

where $\tilde{g}, h$ are the coupling constants for ξ to κ and ψ respectively.

1.10.1 Inflatonic Dark Matter

Now if the decay to fermionic fields ψ is absent and assuming no symmetry breaking in this field, we see that the only way ξ can decay is through the term $\frac{1}{2}\tilde{g}^2\xi^2\kappa^2$, i.e. it requires that two ξ particles interact at close range. Thus the rate of decays will, as in standard nucleosynthesis, be proportional to the concentration of ξ particles in a given region. This leads to an interesting possibility, namely that a large proportion of the ξ particles, which are slowly moving, massive scalar bosons, do not decay because the decay rate falls quickly below that of the expansion rate.

In this case we have an excellent candidate for the dark matter [26] of the universe, which has the following properties:

• It behaves on large scales, like a cosmological constant, $\mu = -p$, the energy density simply being, to zero order, the value $\frac{1}{2}V(\phi_0)$ of the minimum of the effective potential, which is essentially arbitrary (the cosmological constant problem).

• On small scales the equation of state varies, with gravitational collapse induced peculiar velocities, i.e. $\dot{\phi} \neq 0$ and hence $\mu \neq -p$.

Large cosmological constants ($\Omega_\Lambda \sim O(1)$) are coming back into fashion as they explain a number of observations rather well, particularly related to Lyman-α forest, damped systems and the CMB. However, we have here a mechanism which is more flexible than a cosmological constant since on small scales, where there are gravitationally induced peculiar velocities, the equation of state deviates from $p = -\mu$ because of velocity terms $\dot{\xi}$.

The other beauty of this formalism is its simplicity - there is no need for supersymmetric particles or other theories to provide the cold dark matter - the same formalism that produced the quantum perturbations and lead to the large entropy production is also the origin of the dark matter. This possibility is appealing because of its unifying simplicity.

A side-effect of this simplicity is that in local overdensities, the concentration of IDM particles would be large and hence decays to other fields would start again. Searches for such decays in other dark matter candidates are under way, and if detected would give direct information about the physics of the early universe, in particular the mechanics of the inflationary process.

1.10.2 What happens to density perturbations ?

Now from a particle physics point of view the universe is well approximated as exactly homogeneous and isotropic. In particular we note that the perturbations in ϕ have not been included in the above discussion. The question then arises:

What happens to the energy density perturbations when parametric preheating occurs ? This has been discussed recently within the Bardeen approach [28], but is complicated by the fact that the evolution equations are singular. Here we start with a thought experiment.

Since the number of particles in each field mode ξ are increasing exponentially, what happens to the perturbations in the energy density due to the ξ field ? We can identify two natural extremes:

(1) All modes are excited roughly equally, so that spatial variations (perturbations) in the occupation numbers and hence energy density, have roughly the same amplitude after preheating as before.

(2) Since mode occupation numbers grow exponentially, so do spatial variations and hence the amplitude of energy density perturbations grows exponentially during preheating.

Which of these two scenarios is correct ? We will present here a simple argument favouring the second, followed by a preliminary analysis of the problem which backs it up.

Consider the inflaton field towards the end of inflation. It is coupled to the ξ field and hence the quantum perturbations in ϕ should, to some extent be encoded in the ξ field. Thus, even if the ξ field has low occupation numbers at this stage, we expect it to have a spectrum of perturbations. However, this is not necessary.

As parametric preheating begins, the production of ξ particles (which is proportional to the density of ϕ particles because the process is a two ϕ decay) occurs preferentially in the peaks in the ϕ field, and hence in the peaks of the energy density field. The production rate will be exponential, $exp(\mu_k(x^i)z)$, with a larger coefficient $\mu_k(x^i)$ at events x^i where the density is higher than in underdensities. Thus although all mode occupation numbers grow exponentially, because the coefficients are different, the perturbations in the ξ field also grow exponentially (the ratio of exponentials is exponential). Hence we expect the perturbations in the energy density of the ξ field to also grow exponentially. In a way this is the other extreme of the argument presented earlier suggesting inflationary fields as dark matter candidates.

Indeed this simple thought-experiment has subsequently been confirmed by detailed analysis [29], which although complicated by the fact that the equation for the Bardeen potential Φ becomes singular periodically for an oscillating scalar field [102], shows that the $k \neq 0$ density perturbation modes do in fact grow through parametric resonance during reheating. The main subtlety in these analyses are related to the singular behaviour of the equation for the Bardeen potential. To demonstrate this within the covariant approach, examine eq. (1.19) for the simplified case where we ignore the expansion, $\Theta = 0$. This simply reduces the complexity of the equation without introducing any spurious behaviour.

Then we have:

$$\mathcal{A} = -\frac{\dot{\gamma}}{\gamma} \tag{1.46}$$

$$\mathcal{B} = 0 \ \text{ if } \ K = 0 \tag{1.47}$$

$K = 0$ choosing flat background spatial sections, and

$$\frac{\dot{\gamma}}{\gamma} = -\frac{2}{\dot{\phi}}V' \tag{1.48}$$

The equation for X then reduces to (after mode expansion):

$$\ddot{X} - (\frac{k^2}{S^2} - \mathcal{B} - \frac{\mathcal{A}^2}{4} - \frac{\mathcal{A}\dot{\mathcal{A}}}{2})X = 0 \tag{1.49}$$

The crucial thing is that for a second order phase transition where reheating is oscillatory, $\dot{\phi} = 0$ periodically, and hence eq. (1.48) diverges periodically. A $0/0$ situation can never occur for standard potentials since V' is generically proportional to some power of ϕ and this is finite when $\dot{\phi} = 0$. Thus we uncover the same singular behaviour as in the Bardeen formalism. Perhaps more worrying, our definition of the four velocity also diverges because of a similar $(\dot{\phi})^{-1}$ factor. How to obtain a non-singular equation in the covariant approach is not obvious at this stage and is left to future work.

1.11 Nonlinear Quantisation

The quantum-gravity epoch is generally believed to be accompanied by nonlinear geometric and topological fluctuations. If this is correct, the traditional use of first order perturbation theory would seem to be of little use in understanding the real universe. However, there is a narrow regime where the semi-classical methods must link into a true theory of quantum gravity. In this region one must talk of topological objects such as sphalerons, skyrmions and instantons, one must also examine the backreaction of nonlinearity and deal in general with non-perturbative aspects of field theory.

In this section I would like to give a brief overview of methods in nonlinear quantum field theory. This is primarily because the main advantage of the formalism that we have developed here, is that it can be extended in principle to nonlinear perturbations in both a covariant and gauge-invariant manner. However, an important distinction should be made here.

When one speaks of nonlinearity these are two different things depending on the speaker. In particle physics one examines non-linear equations of motion, such as the non-linear Schrödinger equation, or non-linear sigma models, which as a result can have topological solutions: solitons etc... which are by construction non-perturbative, with mass $\propto \lambda^{-1}$ where λ is the appropriate nonlinear coupling constant of the theory. They are interested in the backreaction of nonlinearity on particle production, decay etc.. However, they never consider the backreaction of this nonlinear field configuration on spacetime. In a later section we will consider such a situation and its quantisation.

The consideration of the backreaction on spacetime of nonlinear fields brings us to another extremely rich field of research: namely black hole and singularity physics. This is particularly true if one agrees with the views of t'Hooft for example, who suggests interpretation of black holes as fundamental particles in a third quantised quantum gravity. However, there exists a rather unexplored region lying in the region where $\delta\Phi/\Phi \ll 1$ and black hole formation is not copious. Namely the region where the effects of expansion on the soliton are important and the backreaction on spacetime, while not strong enough to produce a singularity, *is* important and is not amenable to perturbative approaches.

Before continuing this discussion one might ask why the use of complicated variables (e.g. Δ) is useful when we want to study nonlinear perturbations. In this case why do we not simply use the basic fluid quantities such as Θ, ρ together with the Bianchi identities as is done in the fully nonlinear silent universe formalism (see chapter 3 of this thesis) ? This is a valid point if we are interested in field configurations that are nonlinear over the whole of spacetime. However, nonlinear partial differential equations are often characterised by localised nonlinear solutions (non-topological solitons) whose locality is their key feature. In Newtonian gravity, solitonic solutions have been found and it is likely that similar solutions exist in General Relativity. Thus in a cosmological context we would be interested in localised solitons on a FLRW background. Hence we would still be interested in describing the mean behaviour of the universe as nearly FLRW but on small scales one would have nonlinear, localised structures, somewhat like the present cold universe. This brings in the issue of averaging, yet another unsolved problem.

Thus there are reasons for considering the covariant variables in the nonlinear regime although a perturbative expansion becomes dubious. Either way, semi-classical quantisation requires a classical solution to be found and this is difficult since we have coupled nonlinear equations.

1.11.1 Backreaction and the O(N) vector model

In our previous discussion of preheating we considered an effective potential of the quadratic form, so that the resulting equation of motion (eom) was linear. Another very popular model is the $\lambda\phi^4/4$ potential which yields a cubic nonlinearity in the eom, with coupling λ. Thus at finite λ, when ϕ becomes large, standard perturbative approaches fail and the effects of the backreaction can be crucial.

This has been a subject of intense work recently within the paradigm of preheating [26] where there is a huge excitation of inhomogeneous modes & corresponding exponential particle production. To include the effects of backreaction, we discuss the application of the O(N) vector model in the large N limit (see e.g. [24]). This is a non-perturbative approximation which conserves energy and satisfies the Ward identities of the underlying $O(N)$ symmetry. As Boyanovsky *et al* [24] show, the backreaction is very important in stopping the exponential growth of particle number, being even more important than the expansion of the universe. As discussed above however, they are unable to consider the backreaction on the geometry, instead setting the problem only in Minkowski spacetime.

The O(N) vector model is very interesting however, with a perturbative expansion in the number of "pions", N, and not in the amplitude of the field. In this sense it is rather similar to the Zel'dovich solution of structure formation where one considers perturbations in the gravitational potential, thus allowing the study of non-perturbative aspects of structure formation (e.g. caustic & pancake formation). The same situation occurs in gravitational lensing where the perturbation of the light geodesics rather than the intensity of light allows the study of nonlinear intensity fluctuations (see chapters 6-8).

The Lagrangian density and potential for the O(N) vector model are given by:

$$\mathcal{L} = \frac{1}{2}\partial_\mu \phi \partial^\mu \phi - V(\phi \cdot \phi) \tag{1.50}$$

and

$$V(\sigma, \phi) = \frac{1}{2}m^2 \phi \cdot \phi + \frac{\lambda}{8N}(\phi \cdot \phi)^2 \tag{1.51}$$

Here ϕ is an N dimensional O(N) vector, $\phi = (\sigma, \pi)$ and (π) represents $N-1$ "pions". Notice the factor $\lambda/8N$ in the potential, which vanishes in the limit $N \to \infty$ at constant coupling, and hence reduces the problem to the solvable linear case.

To do a consistent treatment one must have the non-equilibrium Green's functions which involves a path integral along a complex contour in time and the doubling of the number of fields corresponding to including backwards and forwards time evolution. The reader is referred to [24] (and references therein) for the technical details, including issues of renormalisation.

One of the very interesting results is that the energy density and pressure for the sigma field σ, are changed from that of a classical scalar field. The expectation value of the (bare) energy density is:

$$\mu = \frac{1}{2}\dot{\phi}^2 + \frac{m^2}{2}\phi^2 + \frac{1}{8\pi^2}\int k^2 dk \left(|\dot{\varphi}_k(t)|^2 + \omega_k^2(t)|\varphi_k(t)|^2\right) - \frac{\lambda}{8}\langle\psi^2(t)\rangle^2 \tag{1.52}$$

while the pressure is given by:

$$p = \dot{\phi}^2 + \frac{1}{4\pi^2}\int k^2 dk \left(|\dot{\varphi}_k(t)|^2 + \frac{k^2}{3}|\varphi_k(t)|^2\right) - \mu \tag{1.53}$$

where $\pi(\mathbf{x}, t) = \psi(\mathbf{x}, t)(1, 1, 1, 1, ..., 1)$, ψ representing the amplitude of the pion fields which in this case are Goldstone bosons if the potential has a global minimum at $\phi \neq 0$. φ_k is defined as

$$\varphi_k(t) = \frac{V_k(t)}{\sqrt{W_k}} \tag{1.54}$$

and the $V_k(t)$ are the mode functions in the creation - annihilation operator expansion of the field $\psi(\mathbf{x}, t)$ (c.f. eq. 1.30). One of the most important results of this is that the effective potential loses its usefulness as far as thermodynamic issues are concerned (it no longer appears in the energy density and pressure).

The mode functions, $\varphi_k(t)$ satisfy, not the Mathieu equation, but the Lamé equation:

$$\left[\frac{d^2}{d\tau^2} + q^2 + 1 + q_0^2 - q_0^2 \mathrm{sn}^2\left(\tau\sqrt{1+q_0^2}, k\right)\right]\varphi_q(\tau) = 0 \tag{1.55}$$

where sn(.,.) is the elliptic Jacobi sine function. The vital feature of this equation is that it has only a single resonance band, unlike the infinite hierarchy that exist in the case of the Mathieu equation. Hence exponential production of high momentum particles is very suppressed by the nonlinear backreaction. This is a case where a more complete analysis actually ***simplifies*** numerical study, since now only low momentum (small k) modes need be followed numerically, while previously all modes could potentially be exponentially amplified. However, the adaptation of the O(N) model to curved, expanding spacetime is yet to be done, but promises exciting results though they may be difficult to obtain, since the $1/N$ nature of the expansion controls the strength of the non-gravitational coupling. The gravitational backreaction however is sensitive to the field amplitude which is nonlinear. The question remains, is there a corresponding way to treat the gravitational backreaction ?

1.11.2 Quantisation of static solitons on Minkowski space

The simplest, illustrative model with nonlinear spatial inhomogeneity one can take is to quantise static solutions in two dimensions on a flat, non-expanding background with a weak coupling to the nonlinearity. This is a long way from being realistic, but we must start somewhere. Further it presents a toy-model for extension to non-flat, expanding geometries. In particular we will briefly discuss the semi-classical quantisation of the kink soliton of e.g. the ϕ^4 Klein-Gordon equation [9].

We consider a scalar field $\phi(x,t)$ in 1+1 dimensions with a Lagrangian:

$$L = \int dx \left[\frac{1}{2}(\dot{\phi})^2 - \frac{1}{2}(\nabla\phi)^2 + \frac{1}{2}m^2\phi^2 - \frac{\lambda}{4}\phi^4 - \frac{m^4}{4\lambda}\right] \tag{1.56}$$

where in the 1+1 case, $\nabla\phi$ is just the first derivative w.r.t. x. This equation is known to have a kink solution - i.e. a topological soliton with different boundary conditions at $x = -\infty$ and $x = \infty$ (hence the name topological soliton since it requires infinite energy to convert the system to one with the same boundary conditions at $x = \pm\infty$). The potential for this system is:

$$V(\phi) = \int dx \left[\frac{1}{2}(\nabla\phi)^2 + \frac{\lambda}{4}\left(\phi^2 - \frac{m^2}{\lambda}\right)^2\right] \tag{1.57}$$

Note that we write $V(\phi)$ even though the potential is a nonlinear *functional* of ϕ, often written as $V[\phi]$. The classical solutions obey the equation:

$$\frac{\partial V(\phi)}{\partial\phi(x)} = -\nabla^2\phi - m^2\phi + \lambda\phi^3 = 0 \tag{1.58}$$

This equation has two static solutions [9]: one constant (the vacuum) and one kink solution $\phi_K(x,t) = \pm(m/\sqrt{\lambda})\ \tanh(mx/\sqrt{2})$.

We will consider here the quantisation of the kink solution and its excitations only. In this case the energy is:

$$V(\phi_K) = (2\sqrt{2}/3)(m^3/\lambda) \tag{1.59}$$

The key step that we will take is to Taylor expand the potential (eq. 1.57) *about the solution* $\phi_k(x,t)$. Now we know that ϕ_K is an extremum of $V(\phi)$ so that the term linear in ϕ will be absent from our expansion. In fact one obtains:

$$V(\phi) = V(\phi_K) + \frac{1}{2}\int dx\eta(x)\left(-\frac{\partial^2}{\partial x^2} - m^2 + 3\lambda\phi_K^2\right)\eta(x) + \lambda\int dx(\phi_K\eta^3(x) + \frac{1}{4}\eta^4(x)) \tag{1.60}$$

where $\eta(x) \equiv \phi(x) - \phi_K(x)$. Now the eigenvalues of the second derivatives of $V(\phi)$, ω_n, evaluated at $\phi_K(x)$ are given by the equation:

$$(-\frac{\partial^2}{\partial x^2} - m^2 + 3\lambda\phi_K^2)\eta_n(x) = \omega_n^2\eta_n(x) \tag{1.61}$$

By changing independent variables to $z = mx/\sqrt{2}$ the equation takes on a Schroedinger-like form:

$$(-\frac{1}{2}\frac{\partial^2}{\partial z^2} + 3\tanh^2 z - 1)\eta_n(z) = \frac{\omega_n^2}{m^2}\eta_n(z) \tag{1.62}$$

which is soluble when placed in a box of length L with periodic boundary conditions [10] and gives a mixed spectrum:

$$\omega_0^2 = 0 \quad \text{with} \quad \eta_0(z) = (cosh^2 z)^{-1} \tag{1.63}$$

$$\omega_1^2 = \frac{3}{2}m^2 \quad \text{with} \quad \eta_1(z) = sinhz/cosh^2 z \tag{1.64}$$

$$\tag{1.65}$$

and

$$\omega_q^2 = m^2(\frac{1}{2}q^2 + 2) \tag{1.66}$$

is the "ladder" part of the spectrum with $q = 0, 1, 2...$ and eigenfunctions:

$$\eta_q(z) = e^{iqz}(3\tanh^2 z - 1 - q^2 - 3iq\tanh z) \tag{1.67}$$

These eigenfunctions have asymptotic values:

$$\eta_q(z) \longrightarrow_{z\to\pm\infty} exp[i(qz \pm \frac{1}{2}\delta(q))] \tag{1.68}$$

where $\delta(q)$ is just the phase shift of the scattering states of the associated Schroedinger problem defined by equation 1.62. When the limit $L \to \infty$ is taken, the spectrum merges into a continuum as usual, and using these normal mode eigenfunctions we may diagonalise the potential (eq. 1.60) and also the Lagrangian near $\phi_K(x)$, to zero order in λ. Now we can use these eigenfunctions to expand $\eta(z)$:

$$\eta(x,t) = \sum_n c_n(t)\eta_n(x) \tag{1.69}$$

and the Lagrangian can be rewritten in a form that is not only diagonal, but is that of a set of harmonic oscillators, one for each eigenfunction (with corrections in λ that we neglect). Thus in quantum field theory one constructs an approximate set of harmonic oscillator states near the classical, static solution $\phi_K(x)$ by quantising the time dependent coefficients $c_n(t)$, which yields discrete energy levels:

$$E_{N_n} = V(\phi_K) + \hbar\Sigma_{n=0}^{\infty}(N_n + \frac{1}{2})\omega_n + O(\lambda) \tag{1.70}$$

i.e. a discrete harmonic oscillator spectrum on top of the classical energy $V(\phi_K)$. Note however that the oscillator analogy breaks down for $n = 0$ since $\omega_0 = 0$. However, we may give the following physical description of the states: the lowest energy $N_n = 0$ state is identified as the quantum kink particle [8] at rest. This is the lowest energy that such a quanta may have, but usually, and certainly in this case, does not correspond to the true vacuum of the theory.

The $n = 1$ state (with $N_n \geq 1$) corresponds to excited states of the kink particle, in analogy with the quantum mechanics of the atom. However, the states with $n \geq 2$, which have a qualitatively different spectrum (given by eq. 1.66), are interpreted [9] as scattering states of the mesons of this theory off the kink particle. The phase shift $\delta(q)$ introduced earlier is just the scattering phase shift for a meson interacting with the kink. The states with $N_n \geq 1$ then correspond to multiple meson scatterings.

Note that, although we have obtained a great deal of information from this relatively simple quantisation procedure, we will not predict any new nonlinear features that are not present in the classical theory. Perhaps more importantly, this was for static solutions in flat, static spacetime. In the expanding case we are faced with the subtleties discussed in earlier sections. However, this second problem must be overcome before the main use of the covariant approach - i.e. extension to weak and strongly nonlinear systems - can be utilised.

What can we expect when we go to an expanding or non-flat background ? Certainly we expect the existence of Hawking radiation. But what will the spectrum look like ? Will geometric solitons be induced, e.g. in the form of gravitational wave solitons ? These are, unfortunately, unsolved questions at present.

1.12 Conclusions

The semi-classical approach to quantisation of cosmological perturbations, although widely applied in the inflationary scenario, has many critics. The present work does not attempt to address these at this stage, taking much from the formalisms developed in the earlier literature. In particular, there is a growing body of work which suggests that the superposition principle [5] is lost in curved spacetimes. This would imply that the construction of the Fock spaces necessary for the (linear) canonical quantisation procedure, is at best locally valid. Thus although using much of the good work developed previously, one of the main aims of the covariant approach is to extend cosmological perturbation theory into the quasi-nonlinear and fully nonlinear, regimes.

[8] The use of the term "particle" here may seem strange, but we use it to describe a discrete energy configuration which has localised form factors, i.e. it is localised in space.

The availability of the fully nonlinear equations for the evolution of the GIC variables thus offers an exciting opportunity. The quantisation of a non-linear equation is technically more challenging, relying on semi-classical quantisation of exact classical solutions such as solitons [9]. However it is also possible that new techniques can be applied from the beautiful new results from duality in string theory (see e.g. Polchinski [45]) with the concomitant development of string soliton and d-brane theory. Only time will tell.

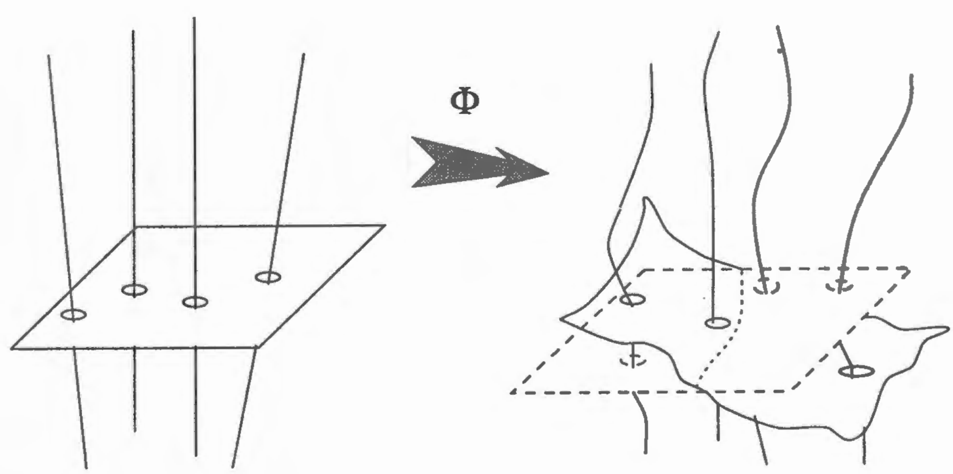

Figure 1.1: The gauge as a map Φ between the worldlines and hypersurfaces of the model background (left) and those of the real geometry.

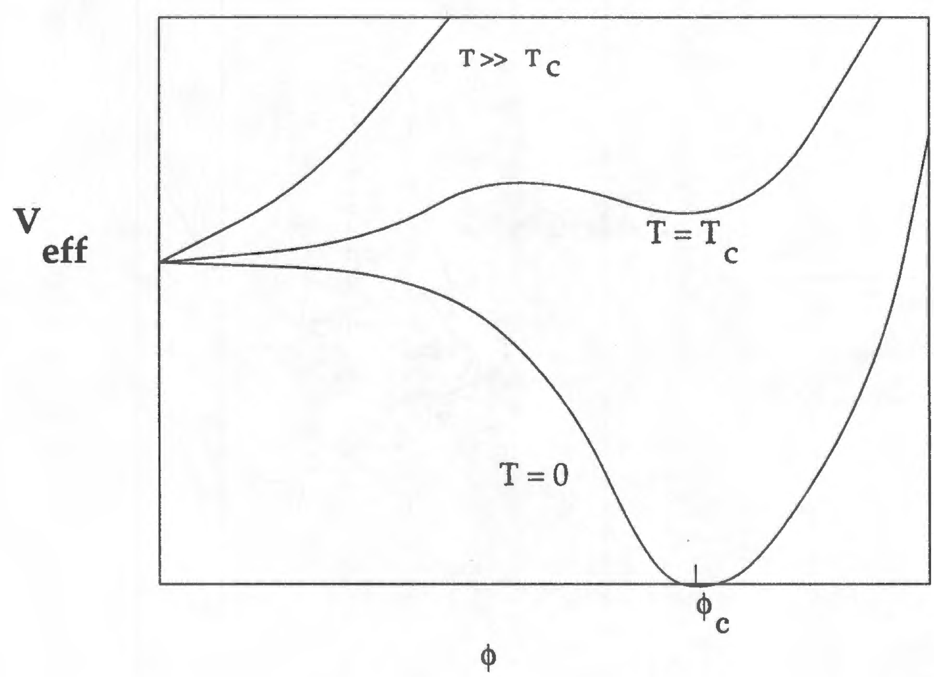

Figure 1.2: The desired behaviour of the finite-temperature effective potential as a function of temperature in the "new-inflationary" model.

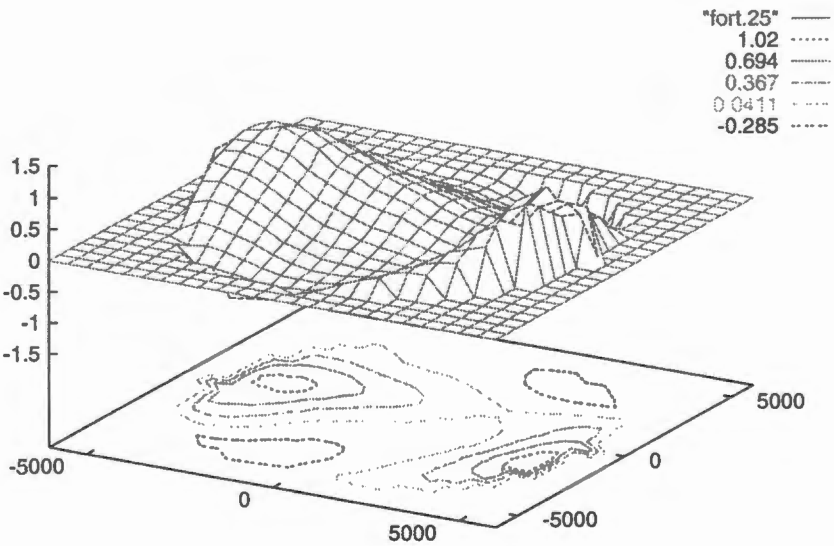

Figure 1.3: This is a plot of the smoothed overdensity, $\delta \equiv \frac{\delta\mu}{\mu}$, within $6000 km/s$ of our position obtained from the IRAS galactic survey using the POTENT method. On the left there is a large overdensity due to the Great Attractor, while on the right there is another overdensity attributed to the Perseus-Pisces cluster. This and the following slices are directly in the supergalactic plane.

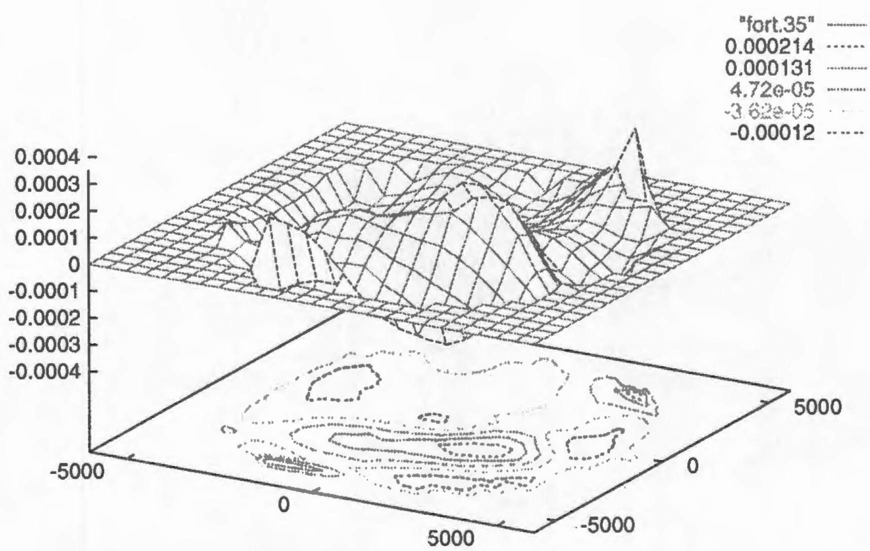

Figure 1.4: A plot of Δ. This and the following figure are to be compared to the previous one for δ. This figure shows the IRAS galaxy data in the covariant perturbation variable Δ used here as fundamental quantisation variable. It is related to ϵ_m and hence δ via eq. 1.13.

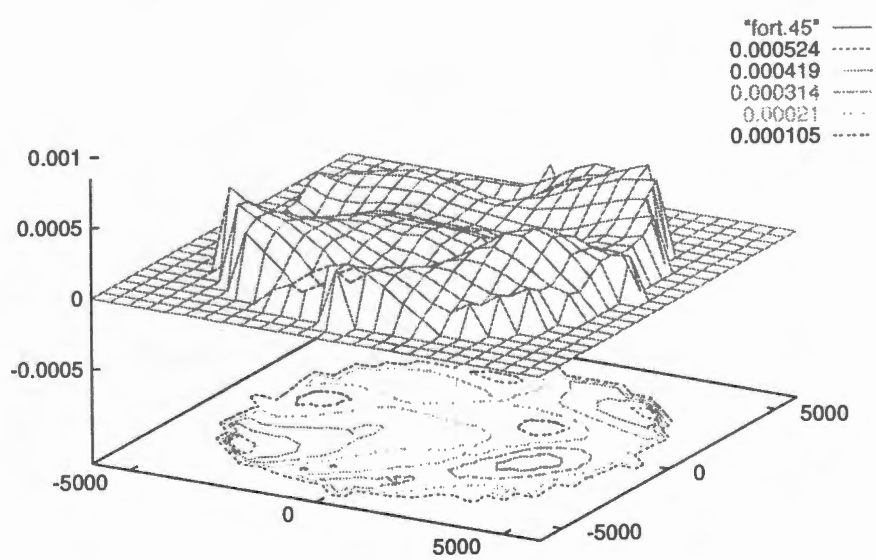

Figure 1.5: This shows the alternative covariant variable describing inhomogeneity not used explicitly here, $\mathcal{D}$, for the same IRAS data. It is always positive, only being zero at comoving density maxima or minima. It was investigated in a classical context in [20].

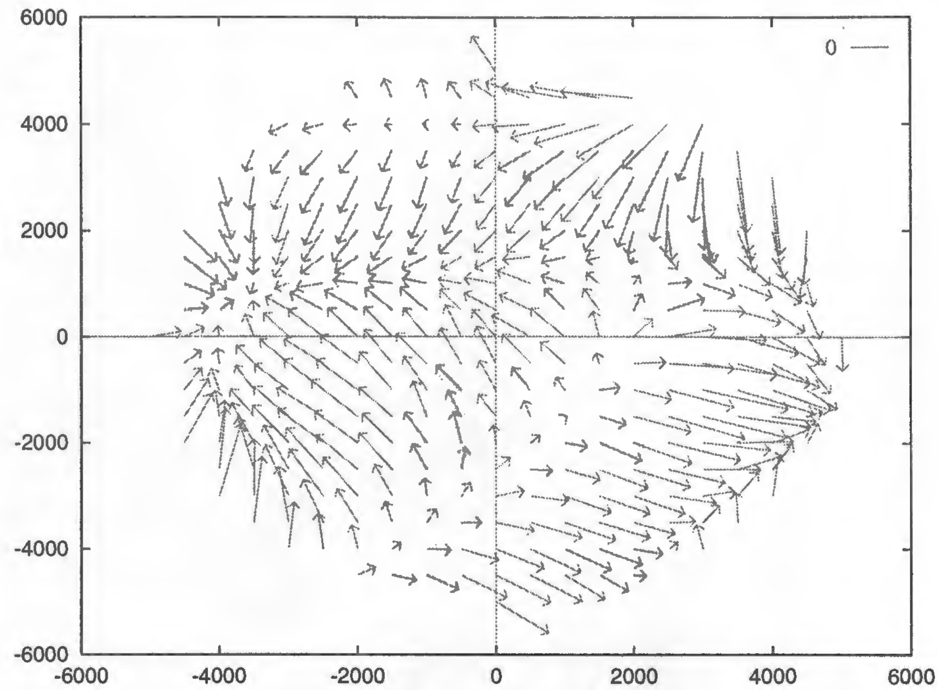

Figure 1.6: The vector field $\mathcal{D}_a$ in the supergalactic plane shows the direction of local infall of the luminous matter in our neighbourhood out to $6000km/s$. The Great attractor and Perseus-Pisces can be seen as points where the density gradient field vanishes on the left and right of the figure respectively.

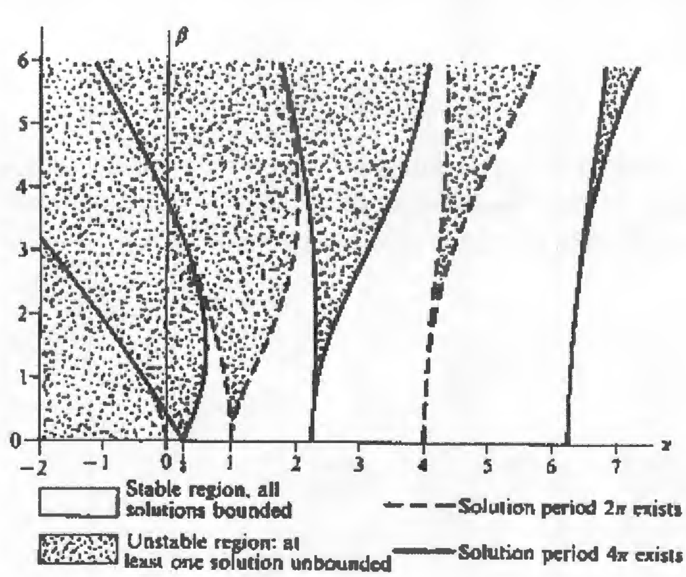

Figure 1.7: The stability chart for Mathieu's equation, β vs. α. The line $\alpha = -\beta$ denoting the boundary of physical wavelengths ($k^2 \geq 0$) is not shown. Figure from [7].

Chapter 2

Covariant Characterisation of Gravitational Waves

"*Nature:* The closer you listen, the more likely you are to hear,
The more languages you speak, the more likely to understand."
Anon.

2.1 Introduction

In this chapter I would like to examine the covariant description of gravitational waves, and compare their properties with the standard, metric-based results. The results are closely related to the work presented in [50].

Both the covariant definition of gravitational waves and the dual question of the physical interpretation of the electric and magnetic parts of the Weyl tensor are subjects of significant debate of recent [73]. In particular the question arises whether one can neglect the magnetic part, H_{ab}, of the Weyl tensor in the Newtonian limit, and whether this is equivalent to neglecting gravitational waves. It is our aim in this chapter to extend this debate and compare the covariant approach to the study of gravitational waves to the standard metric-based approach. By contrast to the covariant approach, the standard metric perturbation approach starting either from the variation of the Einstein-Hilbert action after expansion to second order in the tensor perturbations, or from a direct linearisation of the field equations, yields second order propagation equations. Archetypically, the Bardeen formalism gives the equation of motion:

$$\ddot{H}_T^{(2)} + 2\frac{\dot{S}}{S}\dot{H}_T^{(2)} + (k^2 + 2K)H_T^{(2)} = 0 \tag{2.1}$$

for the gauge-invariant (at first order) amplitude of the tensor perturbation, $H_T^{(2)}$ in the absence of anisotropic stresses [15]. The full tensor metric tensor perturbation is $\sum_k H_T^{(2)} {Q^k}_{ab}$, where the ${Q^k}_{ab}$ are eigenfunctions (polarisation tensors) of the tensor Helmholtz equation:

$$\nabla^2 {Q^k}_{ab} = -\frac{k^2}{S^2} {Q^k}_{ab} \tag{2.2}$$

on the background FLRW spatial sections, and has only two degrees of freedom after the imposition of the transverse ($Q^{ab}{}_{;b} = 0$) and traceless ($Q^a{}_a = 0$) conditions.

2.2 Metric approaches to gravitational waves

Here we will present a brief overview of the basic, flatspace approach to gravitational radiation. Assume that the observer is far from the source in a vacuum background. We expand the metric as: [1]

$$g_{ab} = \eta_{ab} + h_{ab} \tag{2.3}$$

and working in the Lorentz gauge one finds the equations of motion:

$$\Delta_\eta h_{ab} = 0 \tag{2.4}$$

where $\Delta_\eta = \eta^{ab}\partial_a\partial_b$ represents the flat space D'Alembertian operator. When we remove the remaining gauge freedoms by enforcing the h_{ab} to be transverse and traceless, we are left with only two degrees of freedom, identified with the polarisation states of the graviton. There exist plane-wave solutions to this equation, namely:

$$h_{ab} = Q_{ab}e^{ik_r x^r} \tag{2.5}$$

where k_r is the wave vector, $k_r k^r = 0$, and Q_{ab} is the constant polarisation tensor, with $Q^b{}_b = 0$ and $Q^{ab}k_b = 0$. This means that in the approximation that the observer is far from the source, $R_{ab} = 0$ and there is no "Coulombian" contribution to the gravitational field - the Riemann tensor is simply equal to the Weyl tensor:

$$R_{abcd} = C_{abcd} = k_{[c}Q_{d][b}k_{a]} \tag{2.6}$$

from which it follows that

$$R_{abcd}k^d = 0 \tag{2.7}$$

and the Riemann tensor is of Petrov type N, characteristic of a purely radiative solution.

How are the metric perturbations h_{ab} related to the electric, E_{ab} and magnetic, H_{ab}, parts of the Weyl tensor ? Well, to first order in H_{ab}, E_{ab}, the geodesic deviation equation becomes (for a gravitational wave propagating in the x^1 direction):

$$\ddot{\xi}^\alpha = E^\alpha{}_\beta \xi^\beta \quad , \quad \alpha, \beta = 2, 3 \tag{2.8}$$

where ξ^α is the connecting vector between orthonormal tetrads associated with the congruence of geodesics. This means that we can directly attribute the physical effects of linear gravitational waves to the electric part of the Weyl tensor. Of equal interest is the fact that:

$$E_{33} = -E_{22} = -\frac{1}{2}\frac{\partial^2 h_{22}}{\partial x^{0^2}} \quad , \quad E_{23} = -\frac{1}{2}\frac{\partial^2 h_{23}}{\partial x^{0^2}} \tag{2.9}$$

This is of particular relevance when one discusses gravitational waves propagating through boundary layers in the cosmological context (see section 2.7).

[1] Here h_{ab} represents an arbitrary gauge-dependent metric perturbation and should not be confused with the projection tensor used in the covariant approach where $h_{ab} = g_{ab} + u_a u_b$.

The usual treatment of gravitational waves in the *cosmological* context uses the Bardeen metric perturbations h_{ij} which are GI to first order under infinitesimal gauge transformations and are transverse ($h^{ij}{}_{;j} = 0$) and traceless ($h^i_i = 0$), leaving only two independent components. The dynamics can then be reduced to that of a pair of minimally coupled scalar fields, $\phi_{\times,+}$, [27] via the decomposition:

$$h^i_j = h_+ \mathbf{e}_+ + h_\times \mathbf{e}_\times \tag{2.10}$$

where

$$h_{\times,+} = \sqrt{16\pi G}\phi_{\times,+} \tag{2.11}$$

and $\mathbf{e}_+ = \hat{\mathbf{e}}_x \otimes \hat{\mathbf{e}}_y - \hat{\mathbf{e}}_y \otimes \hat{\mathbf{e}}_x$ and $\mathbf{e}_\times = \hat{\mathbf{e}}_x \otimes \hat{\mathbf{e}}_y + \hat{\mathbf{e}}_y \otimes \hat{\mathbf{e}}_x$ are the anti-symmetric and symmetric polarisation tensors for the graviton modes. Quantisation then follows through as outlined in chapter (1).

2.3 The covariant approach to gravitational waves

Within the covariant approach the study of tensor perturbations was first considered by Hawking (1966) [32]. He used the electric part of the Weyl tensor, E_{ab} to characterise them. Later, the magnetic part of the Weyl tensor, H_{ab} was suggested as a better choice [34], partly because there is no Newtonian analogue for it, believed to be a result of the instantaneous propagation of the gravitational force in Newtonian theory. The results presented in this section, together with those of Ellis and Hogan (1996) [47] suggest that, just as in the electromagnetic case, both the electric and magnetic Weyl parts are needed for wave solutions.

E_{ab} and H_{ab} are defined analogously to their electromagnetic counterparts, i.e. they are contractions with the appropriate field strength ($F_{\mu\nu}$ playing the rôle of the Weyl tensor in the EM case):

$$E_{ac} = C_{abcd}u^b u^d \tag{2.12}$$

and

$$H_{ac} = \frac{1}{2}{}^*C_{ghcd}u^b u^d \tag{2.13}$$

where u^c is the four velocity of the fluid and the $*$ represents the Hodge (dual) operator (i.e. contraction with completely anti-symmetric volume element $\eta_{ab}{}^{gh}$). The fully *nonlinear* evolution equations for E_{ab} and H_{ab}, with perfect fluid source, are given by [8]:

$$\begin{aligned} h^m_a h^t_c \dot{E}^{ac} + h^{(m}_a \eta^{r)tsd} u_r H^a_{s;d} &- 2H^{(t}_q \eta^{m)bpq} u_b \dot{u}_p + h^{mt}(\sigma^{ab}E_{ab}) \\ &+ \Theta E^{mt} - 3E^{(m}_s \sigma^{t)s} - E^{(m}_s \omega^{t)s} = -\frac{1}{2}(\mu + p)\sigma^{tm} \end{aligned} \tag{2.14}$$

$$\begin{aligned} h^{ma} h^{tc} \dot{H}_{ac} - h^{(m}_a \eta^{r)tsd} u_r E^a_{s;d} &+ 2E^{(t}_q \eta^{m)bpq} u_b \dot{u}_p + h^{mt}(\sigma^{ab}H_{ab}) \\ &+ \Theta H^{mt} - 3H^{(m}_s \sigma^{t)s} - H^{(m}_s \omega^{t)s} = 0 \end{aligned} \tag{2.15}$$

Notice that the only difference between these equations is in the sign of the second and third terms and the shear source term coupled to the energy density and pressure in the $\dot{E}_{ab}$ equation. Once linearised about a FLRW background, these equations become [50]:

$$\dot{E}_{ab} + \Theta E_{ab} + h^f_{(a}\eta_{b)cde}u^c\nabla^e H^d_f + \frac{1}{2}(\mu + p)\sigma_{ab} = 0 \tag{2.16}$$

and

$$\dot{H}_{ab} + \Theta H_{ab} - h^f_{(a}\eta_{b)cde}u^c\nabla^e E^d_f = 0 \tag{2.17}$$

It is the "curl" terms $h^f_{(a}\eta_{b)cde}u^c\nabla^e$ that yield the travelling gravitational waves, in analogy with the propagation of electromagnetic waves.

2.4 Closed evolution equations for linear gravitational waves

A crucial rôle is played by the divergence constraints:

$$h^t{}_a E^{as}{}_{;d}h^d{}_s - \eta^{tbpq}u_b\sigma^d{}_p H_{qd} + 3H^t{}_s\omega^s = \frac{1}{3}h^{tb}\mu_{;b}\,, \tag{2.18}$$

$$h^t{}_a H^{as}{}_{;d}h^d{}_s + \eta^{tbpq}u_b\sigma^d{}_p E_{qd} - 3E^t{}_s\omega^s = (\mu + p)\omega^t\,, \tag{2.19}$$

which again are fully nonlinear. Linear tensor perturbations alone can be chosen by imposing the restrictions that both the linearised divergences of E_{ab} and H_{ab} vanish (the other terms being $2nd$ order). Then eq's (2.18,2.19) imply:

$$\frac{1}{3}h^{ak}\mu_{,k} = 0 \tag{2.20}$$

$$(\mu + p)\omega^a = 0 \tag{2.21}$$

These conditions are the analogue of the transverse condition on tensor perturbations in the metric approach, and come from expanding equations (2.18, 2.19) to first order about a FLRW background.

By differentiating eq's (2.16,2.17) and using the linearised shear evolution equation:

$$\dot{\sigma}_{ab} =^{(3)} \nabla_{(a}a_{b)} - \frac{2}{3}\Theta\sigma_{ab} - E_{ab} \tag{2.22}$$

For purely tensor perturbations this contains only the last two terms, since there is no acceleration since there are no pressure gradients. By substitution we thus replace $\dot{\sigma}_{ab}$ in terms of σ_{ab} and E_{ab}.

Note that the belief [49] that $H^{ab}{}_{;a} = 0 \;\Rightarrow H^{ab} = 0$ has, fortunately for the covariant approach, been shown to be incorrect [51], so that a linear analysis of gravitational waves in the covariant approach is fully consistent.

In addition, notice that since the Weyl tensor is the trace-free part of the Riemann tensor, both E_{ab} and H_{ab} are trace-free, again like the tensor perturbations of the metric-based approach. Once these conditions have been imposed, one might suspect that both E_{ab} and H_{ab} would describe precisely the same things, i.e. that the evolution equations would be invariant under the transformation $E_{ab} \leftrightarrow H_{ab}$. However, we will show that this is not the case in general. In fact, it is only true when $\mu + p = 0$, i.e. the background spacetime is vacuum (Milne or Minkowski) or de Sitter.

In general, the linearised equations for E_{ab} and H_{ab} are not even of the same *order*, the former having a third order equation (after eigenfunction expansion) and the latter a second order one. In full generality we have, after tensor eigenfunction expansion with $E_{ab} = \sum E_k Q^k{}_{ab}$ [50]:

$$\dddot{E} + \left[3\Theta + \frac{\dot{B}}{B}\right]\ddot{E} + \left[\frac{7\Theta^2}{9} - \frac{7}{6}(\mu + 3p) + A - \frac{7\Theta}{3}\frac{\dot{B}}{B}\right]\dot{E} + \left[\dot{A} - A\frac{\dot{B}}{B} + \frac{2}{3}\Theta A - B\right]E = 0 \tag{2.23}$$

where

$$A = \frac{2}{2}\Theta^2 - 2p + \frac{k^2}{S^2} \tag{2.24}$$

and

$$B = -\frac{1}{6}(\mu + p)\left[3\Theta(1 + 3\frac{p}{\mu}) - 3\frac{\dot{p}\mu - \dot{\mu}p}{\mu(\mu + p)}\right] \tag{2.25}$$

where we have assumed that $(\mu + p) \neq 0$. If this is not true then the equation reduces to second order immediately. In contrast, the H_{ab} equation is relatively simple after eigenfunction expansion:

$$\ddot{H} + \frac{7}{3}\Theta\dot{H} + \left[\dot{\Theta} + \Theta^2 + \frac{1}{2}(\mu - p) + \frac{k^2}{S^2}\right]H = 0 \tag{2.26}$$

Note that the only difference between the coefficient of H in the above equation and A (given by eq. 2.24) is the replacement $\mu - p \to \mu + p$. The reason H_{ab} has a second order equation is for the following essential reason: the following constraint equation (with $\omega_{ab} = 0$) [8]:

$$H_{ab} = -\frac{1}{2}h^k_{(a}h^l_{b)}\sigma_{km;n}\eta_l{}^{omn}u_o \tag{2.27}$$

allows the covariant curl of the shear to be replaced by H_{ab}. In the case of the $\ddot{E}_{ab}$ equation, there is a σ_{ab} term which cannot be removed without differentiating again (since there is no constraint which relates σ_{ab} to E_{ab}).

Note that if we have non-perfect fluid matter, (or multi-fluids) so that there is an anisotropic stress, then there will be additional source terms in the equations so that one cannot even close the equations at third order [8].

2.5 Discussion of the nature of the equations

The appearance of a third order equation for E is very surprising. First of all, the standard theory discussed in section (2.2) gives a second order evolution equation, and secondly, force laws are generally expected to be formulated as second order evolution equations.

In fact there is a very interesting parallel with the situation in topologically massive gravity in $2+1$ dimensions. In the case where the Lagrangian is just the Einstein-Hilbert one: $\mathcal{L} = \sqrt{g}R$, there is no dynamics since the Riemann tensor is just proportional to the Ricci tensor - the Weyl tensor being identically zero. Since there is no tidal field it is impossible to have gravitational waves and this is the standard argument against the relevance of lower-dimensional studies of General Relativity. However, once we consider quantisation of this theory a remarkable thing occurs. Due to the celebrated theorem by t'Hooft, any terms which do not break gauge symmetry of the Lagrangian will *always* appear at first loop in a quantum expansion of the theory. Hence they should be included *a priori*. The Chern-Simons term is exactly such a term ! The appropriate Lagrangian is then [37]:

$$\mathcal{L} = \sqrt{g}R + \frac{1}{2\mu}\epsilon^{abc}\Gamma^{d}{}_{be}\left(\partial_a\Gamma^{e}{}_{cd} + \frac{2}{3}\Gamma^{e}{}_{af}\Gamma^{f}{}_{cd}\right) \tag{2.28}$$

Physically this crucial addition has several effects. Firstly it allows gravitational waves. This is partly because the Chern-Simons term is the exact analogue of the Weyl tensor in four dimensions - it is symmetric and traceless with the same symmetries. Secondly, the Chern-Simons term induces a topological mass for the graviton, so that the gravitational force has a finite range in $2+1$ dimensions. This is the exact analogue of the Meissner effect in superconductors where photons gain a mass and hence cannot reach deeper than the surface of the superconductor. Further. the Chern-Simons term allows gravitational radiation - but causes the equation of motion for the metric perturbations to be third order, once the above Lagrangian has been linearised [37].

However, the crucial difference is that the constraints of gravity can be used to reduce the equation of motion of the physical variable to that of the Klein-Gordon equation, with the graviton having a mass $|\mu|$. Thus even this promising candidate for third order dynamics for gravitational radiation fades to second order evolution in the end. But this is not serious when one realises that this example was formulated in Minkowski spacetime without any sources. When source terms are included (such as due to expansion in a nontrivial matter background) the evolution equation cannot be reduced to second order.

Perhaps a final clarification belongs to the rôle of the shear tensor. It obeys a second order evolution equation [50]:

$$\ddot{\sigma}_k + \frac{5}{3}\Theta\dot{\sigma}_k + \left[\frac{k^2}{S^2} + \frac{\Theta^2}{9} + \frac{4\mu}{3} + 3p\right]\sigma_k = 0 \tag{2.29}$$

Once the shear has been determined, one can immediately determine both the magnetic and electric parts of the Weyl tensor. H_{ab} is just the covariant curl of the shear, eq. (2.27), while E_{ab} is determined from the shear evolution equation, (2.22). So the truly fundamental variable is the shear, and it has a second order evolution equation for linear tensor perturbations of any FLRW background.

2.6 Solutions - analytic and numerical

Here we will examine the evolution equations of the magnetic part of the Weyl tensor in some of the simplest cases; namely Minkowski, Milne (Vacuum FLRW), de Sitter and Einstein-de Sitter.

2.6.1 Gravitational waves in vacuum

In Minkowski, the expansion is zero and we have:

$$\Delta_\eta H_{ab} = \Delta_\eta E_{ab} = 0 \tag{2.30}$$

where again Δ_η is the flat-space D'Alembertian. These are the precise analogues of eq. (2.4) except that no gauge conditions are needed.

In an empty expanding universe the equations (2.26,2.23) reduce to:

$$\ddot{H} + \frac{7}{3}\Theta\dot{H} + (\dot{\Theta} + \Theta^2 + \frac{k^2}{S^2})H = 0 \tag{2.31}$$

with the E_{ab} equation having exactly the same form, with the replacement $H \to E$. The appropriate FLRW vacuum background is the Milne solution which has: $\Theta = 3t$. Thus the equation becomes:

$$\ddot{H} + \frac{7}{9t}\dot{H} + (\frac{9k^2 - 2}{9t^2})H = 0 \tag{2.32}$$

This is almost Bessel's equation except for a missing term proportional to H. For small k, this has power law solutions:

$$H(t) = c_1 t^{\alpha_+} + c_2 t^{\alpha_-} \tag{2.33}$$

where

$$\alpha_\pm = -3 \pm \sqrt{12 - k^2}$$

For large k it pocesses oscillatory solutions as we expect for gravitational waves on small scales, while for $k^2 < 12$ one has pure decaying mode evolution. In figures (2.1-2.3) we see that the envelope amplitude oscillates and the period of oscillation increases with time. We can gain further insight into this equation by making the substitution:

$$H(t) = t^\mu f(t) \tag{2.34}$$

To eliminate the $\dot{H}$ term requires $\mu = -\frac{7}{2}$, so that the equation becomes:

$$\ddot{f} + \frac{(k^2 - 11/4)}{t^2} f = 0 \tag{2.35}$$

This is simply the equation for a harmonic oscillator with frequency $\propto t^{-2}$. Thus we expect oscillatory solutions when $(k^2 - 11/4) > 0$. Indeed this is borne out in numerical solutions (see figure 2.1). Thus qualitatively we have $\sin(1/t)$ oscillations with a non-trivial envelope evolution, somewhat similar to the behaviour of the Airey functions. The scaling behaviour of the frequency can be seen in figures (2.1 - 2.3) with the pattern of oscillations essentially invariant under dilations of the time coordinate.

2.6.2 Gravitational waves in de Sitter Space-time

Although the equations for E_{ab} and H_{ab} are the same in de Sitter spacetime, they are more complex than in the vacuum case because of the additional $\frac{1}{2}(\mu - p)$ term. If we consider a cosmological constant Λ. then we have:

$$\mu = \Lambda \tag{2.36}$$

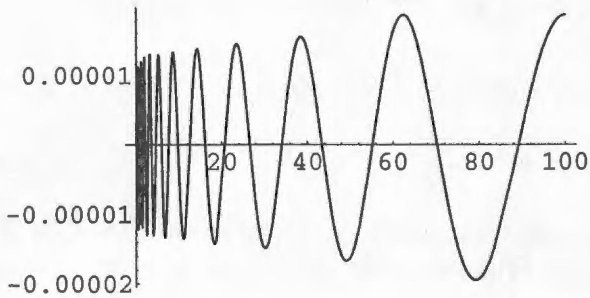

Figure 2.1: The modes of the magnetic part H_{ab}, in the case of vacuum (Milne) solution, for large wavenumbers $k \gg 1$. Initial conditions were $H(0) = 0$ and $\dot{H}(0) = \epsilon$, a small constant. Note that the envelope appears to be growing with time (compare the following two figures).

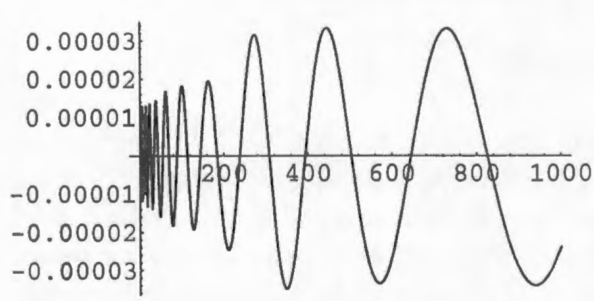

Figure 2.2: As in the previous figure, except now observed over a time frame which is ten times longer. Note that the envelope appears now to be modulated with time. However, the period still grows monotonically.

$$p = -\Lambda \tag{2.37}$$

so that $\frac{1}{2}(\mu - p) = \Lambda$. Thus the equations become:

$$\ddot{H} + \frac{7}{3}\Theta\dot{H} + (\Theta^2 + \frac{k^2}{Ae^{2\Theta t/3}} + \Lambda)H = 0 \tag{2.38}$$

where the specification to de Sitter space gives $\Theta = \sqrt{24G\mu}$ and $\dot{\Theta} = 0 = \dot{\mu}$.

This equation can be converted to Lommel's equation by making the following change of time coordinate:

$$\xi = Ae^{2\Theta t/3} \tag{2.39}$$

The equation becomes:

$$H'' + \frac{9}{2\xi}H' + \left(\frac{k^2}{4\Theta^2}\xi^{-3} + \frac{\Theta^2 + \Lambda}{\xi^2}\right) H = 0 \tag{2.40}$$

with $' = d/d\xi$. This can be solved in terms of Bessel functions. The solution in terms of ξ is:

$$H_k(\xi) = \xi^{-7/4} J_\nu(\beta\xi^{-1/2}) \tag{2.41}$$

where J_ν is the Bessel function of order ν and the parameters are:

$$\nu^2 = 4[(7/4)^2 - \Theta^2 - \Lambda] \tag{2.42}$$

and $\beta^2 = k^2/\Theta^2$.

This is interesting for at least two reasons: firstly the spectrum (i.e. the variation with k), has a $J_\nu(k)$ scaling, i.e. the spectrum depends only on a linear scaling of the argument of the Bessel function.

Secondly, the nature of the solution varies depending on the field to which ν belongs. Depending on the value of μ, which controls the values of Θ and Λ, ν can either be integer, rational, real or pure imaginary. Since the Bessel function changes quite drastically depending on its order, so will the solution H_k.

2.6.3 Einstein-de Sitter universe

The flat, dust, Einstein-de Sitter universe is perhaps the simplest model in which the symmetry between the orders of the equations for E and H is broken. We will not discuss the evolution of the Electric Weyl part here. Now we have $S(t) \propto t^{2/3}$ and the equation (2.26) for the magnetic modes becomes (absorbing the arbitrary scale-factor normalisation into k):

$$\ddot{H} + \frac{14}{3t}\dot{H} + \left[\frac{5}{2t^2} + \frac{k^2}{t^{4/3}}\right] H = 0 \tag{2.43}$$

Again this equation has power law solutions in the long-wavelength limit ($k \to 0$), where the equation is similar to that in the Milne case. The power law solutions are given by:

$$H(t)_{k=0} \propto t^{\alpha_\pm}, \qquad \alpha_\pm = -\frac{11}{6} \pm \frac{\sqrt{31}}{6} < 0 \tag{2.44}$$

Thus, as in the open, expanding vacuum case, there are only decaying tensor modes on superhorizon scales in the presence of matter, for a flat, dust universe.

We plot the evolution of H in figures (2.4-2.5).

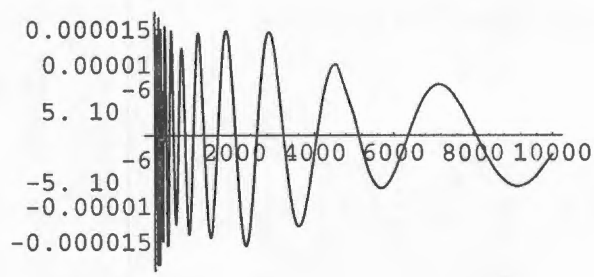

Figure 2.3: The same H modes in a Milne universe, but with a further $\times 10$ increase in the length of time evolution showing the extended oscillatory nature of the envelope.

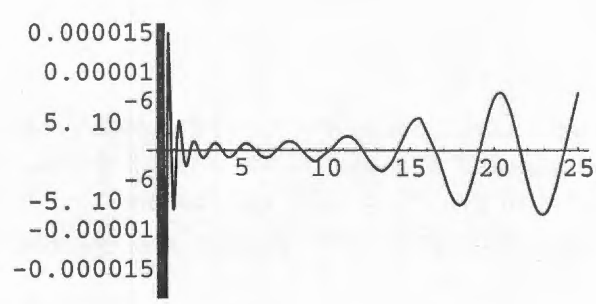

Figure 2.4: The evolution of H_k in an Einstein-de Sitter universe near the time origin for large wavenumbers, $k \gg 1$. Note the rapid decay initially followed by the regeneration of the envelope.

2.7 Sharp phase transitions

If inflation is correct, then there must have been a transition at the end of inflation from the near-de Sitter phase to a FLRW radiation phase, often treated as an instantaneous 3-surface, although in reality this reheating is not instantaneous. The question posed is: "can there be amplification of perturbations across this boundary ?" The standard answer is that density perturbations can be amplified [38], but that tensor perturbations cannot. This is simply a result of the Darmois junction conditions which require that the three-metric and the extrinsic curvature of the three-metric be continuous across the boundary.

The key point for us is that this theorem of no-amplification does not carry through in the covariant case since both the electric and magnetic parts of the Weyl tensor will depend on the second derivatives of the metric perturbations, which are unconstrained by the junction conditions. Thus we cannot agree immediately with the conclusion that gravitational waves cannot be amplified across a sharp phase transition, if we impose Darmois junction conditions across the join.

For the linearised metric, eq. (2.3), the Riemann tensor is:

$$R_{abcd} = \partial_d \partial_{[a} h_{b]c} - \partial_c \partial_{[a} h_{b]d} \tag{2.45}$$

and the electric part of the Weyl tensor in tetrad form can be written:

$$E_{\hat{a}\hat{b}} = -R_{\hat{0}\hat{a}\hat{0}\hat{0}} - \frac{1}{2}R_{\hat{a}\hat{b}} + \frac{1}{6}\delta_{\hat{a}\hat{b}}(R + 3R_{\hat{0}\hat{0}}) \tag{2.46}$$

From this, and equations (2.9), we see the explicit dependence on second derivatives of the metric perturbations, which are unconstrained by the Darmois conditions through a sharp phase transition. In particular we see from our initial discussion on gravitational waves (section 2.2 and equations 2.8, 2.9), that E_{ab} only depends on the second time derivative of the metric perturbations and appears therefore to be unconstrained through a sharp phase transition. Note however, that the divergence of the shear is related to the spatial gradient of the expansion, eq. (3.31) (i.e. the extrinsic curvature) via the $(0, \nu)$ constraint equations. This together with the 'div E' and 'Div H' constraints might be sufficient to enforce continuity of the Weyl tensor through a space-like boundary junction. However, this deserves further careful work.

2.8 Discussion

Here we have discussed the evolution of gravitational waves, mainly in the cosmological context, in terms of the electric and magnetic parts of the Weyl tensor, E_{ab} and H_{ab}.

It is interesting that the evolution equations for E_{ab} and H_{ab} are so different, because in being so it seems to imply a radical break with the analogy with electromagnetism, which hitherto had been thought to carry through almost completely, at least qualitatively. However, Ellis and Hogan [47] have shown that indeed the electromagnetic field behaves very similarly in an expanding FLRW background with a similar breaking in the order of the evolution equations. Why is there this amazing parallel between electromagnetism (which is an Abelian U(1) gauge theory) and gravity (which is a non-Abelian gauge theory) ? It is essentially because of the requirements of Poincaré invariance which are applied to both theories and are very strong.

Just as interesting are the cases in which the evolution equations reduce to the same, second order equation: i.e. in vacuum and in de Sitter spacetime, the usual model for the inflationary background. Similarly, questions naturally arise about the implications of different order evolution equations in the more general context. These are issues for further study.

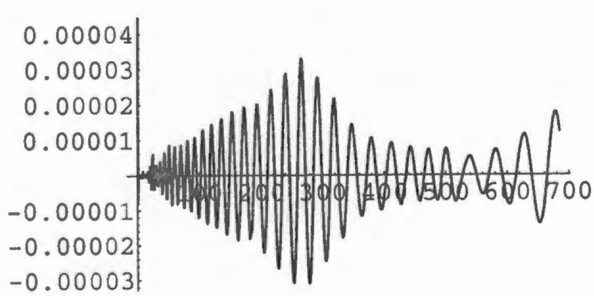

Figure 2.5: The same EdS universe and wavenumber as above, with extended time evolution. Note that the period of primary oscillations grows more slowly than in the open, expanding, vacuum case.

Chapter 3

Silent Universes and Vorticity

"Now entertain conjecture of a time,
When creeping murmur and the poring dark
Fills the wide vessel of the universe"

Henry V, *W. Shakespeare*

3.1 Introduction

In this chapter we review the silent universe formalism, which represents a significant step in the study of nonlinear matter-dominated General Relativity, and consider relaxation of one of the main assumptions, namely the requirement that the velocity field be irrotational. The main aim of doing this is to see whether the reduction of the field equations to ordinary differential equations is preserved and how it affects the issue of integrability of the field equations. Further, since the vorticity has a growing mode in the nonlinear density regime and in fact shares the same caustics as the density field in the Lagrangian frame [69], the use of the traditional silent universe formalism to study the end-points of gravitational collapse of clusters of galaxies may perhaps not show the full picture. In fact, because of the growth of the vorticity in the nonlinear regime, we can see immediately that the subspace of irrotational silent universes is unstable to perturbations. However, it is unknown whether they share the same generic attractors as dust universes with vorticity included. We will show that the silent nature of the flow is not destroyed by including vorticity. We will call such models vortical silent universes (VSU's). This is simply for convenience and brevity and is not a claim that such models have been shown to be integrable, i.e. consistent with the constraint field equations.

A further step is taken by identifying connections with Mixmaster dynamics and Asymptotically Velocity Term Dominated (AVTD) solutions (section 3.9.1).

3.2 Irrotational silent universes

There are several approximation schemes to the ten evolution and constraint equations of General Relativity (GR). There are at least four features of the field equations of GR which make them particularly difficult to deal with:

(1) They are nonlinear,

(2) They are partial differential equations (PDE's), and

(3) They are coupled.

(4) There exist non-linear constraint equations to be satisfied by solutions to the evolution equations.

As far as I am aware, there is no way to eliminate problem (3) in general. The standard approach is to linearise the PDE's using perturbative methods, which eliminates problem (1), but in general still leaves us with problems (2) and (3). Problem (2) is then usually overcome by using expansions into eigenfunctions (e.g. the Fourier transform in flat background), with gradients replaced using the Helmholtz equation. This approach has provided us with a wealth of conceptual understanding of our universe, which has come essentially only because of the near isotropy of the CMB. Had we found ourselves in a universe with a highly anisotropic CMB, with $\Delta T/T \simeq 1$, linear perturbation theory would have been useless and we would have had to study the fully nonlinear equations in much more depth before now. The expansion into eigenfunctions would probably also not work since this only holds in cases where the eigenfunctions of the Helmholtz equation form a complete basis set for functions on the spatial sections, a non-trivial problem. Thus problem (2) would remain unsolved. Similarly problem (3) only disappears when 2nd order terms can be dropped, i.e. in linear theory. Otherwise converting the problem to Fourier space introduces mode-couplings which causes transfer of power from one length scale to another.

The silent universe formalism however, solves problem (2) by converting the Einstein Field Equations (EFE) into a set of coupled ordinary differential equations, (ODE). The formalism is exact, nonlinear, and is obtained via the assumptions of a pressure-free dust medium, a zero magnetic part of the Weyl tensor ($H_{ab} = 0$), and an irrotational velocity field, $\omega_a = 0$. The first assumption ensures that there are no sound waves, the second that there are no gravitational waves, although we should note that the interpretation of H_{ab} is still not fully understood. Physically the idea behind silent universes is that each worldline evolves separately from the rest of the universe, governed only by its initial conditions and local values of the gravitational field, encoded in the density, shear, expansion and E_{ab}.

This locality means that there is no spatial communication between worldlines, and that the PDE's of GR reduce to nonlinear ODE's. This is a remarkable reduction of complexity but, just as with linear perturbation theory, we must know the range of validity of the assumptions that go into the approximation scheme.

Problem (4) is in many ways the most insidious. To check integrability one must ensure that the constraints are preserved under time evolution. This involves repeatedly taking time derivatives of the constraints and substituting from the evolution equations until identities are obtained. Even in relatively simple cases, this process can require four or five time derivatives to be taken [8].

3.2.1 The dust and vanishing H_{ab} assumptions

These are both vital for the reduction of the field equations to ODE's and hence cannot be dropped. However, without pressure one obtains shell crossings in the nonlinear regime and this is obviously not a good description in the latter stages of collapse on small scales since hydrodynamical pressure effects stop e.g. the gravitational collapse of galaxies and afterwards virialisation takes place. However, large clusters and superclusters of galaxies are not yet virialised. Further, it is reasonable to assume that the qualitative behaviour of matter is determined by gravity only, i.e. that the generic collapse to sphere, pancake or spindle configuration is determined by gravity only. Then the dust assumption is reasonably valid. Further, on large scales the matter is very cold and the radiation density is low so that the pressure on average is close to zero.

The vanishing magnetic part of the Weyl tensor assumption, $H_{ab} = 0$, is much more controversial, however. It has been shown [55] that there are consistent solutions with $\omega_a = H_{ab} = 0$ for perfect fluids. However it has also been shown that at second order, scalar perturbations of FLRW models give rise to a non-zero H_{ab}. This mixing is to be expected in general since the consistent splitting into scalar, vector and tensor modes is a linear occurance.

Putting $H_{ab} = 0$ implies that $\dot{H}_{ab} = 0$ (which becomes a new constraint equation) and through the H_{ab} constraint equation [8] and the evolution-turned-constraint equation $\dot{H}_{ab} = 0$ this puts restrictions on the spatial variations of the shear and E_{ab}.

Now even at linear level, the magnetic part of the Weyl tensor can be shown to have both gravitational wave, $H_T^{(2)}$, *and* scalar perturbation, Φ, contributions [35]. Φ is the gauge-invariant Bardeen potential characterising some aspects of gravitational collapse and $H_T^{(2)}$ is, as in the previous chapter, the TT Bardeen tensor perturbation.

The real question is whether there are fully inhomogeneous, integrable models with $H_{ab} = 0$. This is perhaps suggested by the fact that H_{ab} has no Newtonian equivalent (although see the discussions by [72] and [73]). However, such reasoning is very dangerous since the integrability conditions of GR are so restrictive. Indeed, recent work [80] conjectures that there are no fully inhomogeneous (type I) silent models. On the issue of the physicality of the $H_{ab} = 0$ assumption, work by Mutoh *et al* [86] investigated "quiet universes" - linear magnetic perturbations of silent universes - and found them to grow, leaving the silent universes unstable.

3.2.2 The irrotational assumption

The irrotational assumption appears of a different class to the other two, as it is not associated with wave propagation. This is essentially due to the fact that the vorticity follows from a first order conservation law (see eq. 3.20). Thus including vorticity is not expected to allow communication between different world lines, and the irrotational assumption seems simply one of convenience. The effect of vorticity on the issue of integrability, is by contrast, almost completely unknown, and is discussed in section (3.7.3).

Now it is generally believed that the vorticity on large scales is negligible. This follows from the well known fact that in the linear regime and in the absence of dissipative effects, the vorticity has only a decaying mode, proportional to S^{-2}, where S is the scale factor. Further, the popularity of inflation is taken to imply that any vorticity present before the near-de Sitter phase would be hugely diluted leaving almost none afterwards, since scalar fields do not support vorticity. It is true that there will be no vorticity in the inflaton field but not necessarily true of any other non-scalar fields that it was coupled to. Nevertheless, it is likely that on large scales the vorticity can be neglected (see section 3.6 for further discussion). This means that silent universes are stable to vortical perturbations in the linear regime.

However, in the nonlinear regime the vorticity gains a growing mode and in fact it has been shown that in the dust case it diverges with the density at shell-crossings, sharing the same caustics [69]. Thus realistic studies of the endpoints of gravitational collapse of dust spacetimes should include vorticity since silent universes are unstable to vortical perturbations once nonlinear collapse sets in. Thus it may be possible to have a silent spacetime where parts collapse to vortical singularities while the expanding parts are well described by the irrotational silent equations.

3.3 Dynamics of irrotational silent universes

Even though the assumptions $p = \omega_a = H_{ab} = 0$ allow a reduction of the field equations from PDE's to ODE's, one might naively suspect that the corresponding number of ODE's would be equal or larger than the original number of field equations since we need evolution equations for density, expansion, shear and E_{ab}.

However, there is a theorem [55] which states that for perfect fluids with $\omega_a = H_{ab} = 0$, there exists an orthonormal tetrad which is an eigenframe for both the shear and E_{ab}. This means that they can be simultaneously diagonalised. This is a direct result of the "div H" equation which in this case is:

$$\eta^{tbpq} u_b \sigma_p^d E_{qd} = 0 \tag{3.1}$$

which is a statement of the fact that σ and $\mathbf{E}$, taken as matrices, commute with respect to our orthonormal tetrad, i.e. $[\sigma, \mathbf{E}] = 0$.

Further, in all but two special cases, the tetrad vectors are orthogonal to the surfaces defined by the four velocity, u^a, and hence the metric can also be diagonalised. This together with the fact that a dust equation of state implies geodesic flow (vanishing acceleration, $\dot{u}^a = 0$) means that the Einstein evolution equations reduce to just six ordinary differential equations - one each for the density and expansion and two each for the shear and electric part of the Weyl tensor.

3.3.1 Evolution equations

The silent universe formalism is fully covariant and because there is no reference background, has no gauge problems. Because of the vanishing vorticity, the four velocity can be used to define space-like hypersurfaces to which it is orthogonal everywhere. In this case proper time can be used ($\dot{u}^a = 0$) to parametrise events on the worldlines and we have a comoving synchronous coordinate system.

However, one is free to use a tetrad approach instead. This has proven to be very useful in studies of the properties of silent universes ([49] and refs. therein). One may define an orthonormal tetrad $\{u^a, e^a_\mu\}; u_a e^a_\mu = 0, e^a_\mu e^\nu_a = \delta^\nu_\mu$, where Greek indices label the three spacelike tetrad vectors and Latin indices label coordinates. This is where the previous theorem becomes useful since we may choose our tetrad vectors e^a_μ to be aligned with the eigenvectors of σ_{ab} and E_{ab}. Thus we may write [53]:

$$E_{ab} = \sum_{\mu=1}^{3} E_\mu e_{\mu a} e_{\mu b}, \quad \sigma_{ab} = \sum_{\mu=1}^{3} \sigma_\mu e_{\mu a} e_{\mu b} \tag{3.2}$$

with $\sum E_\mu = \sum \sigma_\mu = 0$ and $\sigma^2 = \frac{1}{2} \sum \sigma^2_\mu$, $E^2 = \frac{1}{2} \sum E^2_\mu$. From this we see why only six equations are needed since σ_{ab} and E_{ab} have only two independent components each, the zero "trace" conditions eliminating the third in each case.

In general the metric is also diagonal [55] and in this case we have:

$$u^a = -\delta^4_a \quad , \quad e_{\mu a} = \ell_\mu \delta^\mu_a \tag{3.3}$$

where the ℓ_μ are the scale factors of the metric in the three directions defined by the e, and in the FLRW case coincide with the scale factor S. In general we may define an average expansion via:

$$\Theta = 3\frac{\dot{\ell}}{\ell} \tag{3.4}$$

where $\ell^3 = \ell_1 \ell_2 \ell_3$.

The dynamical variables are thus $\rho, \Theta, \sigma_1, \sigma_2, E_1$ and E_2. Their evolution equations are nonlinear and coupled, but do not depend explicitly on time, i.e. they form an autonomous 6-dimensional system in which the trajectories do not cross or change in time in the full phase space. This is a further simplification which was not implied by the reduction to ODE's and allows the use of standard phase-plane analysis to study evolution and attractors. In fact, General Relativity will always reduce to an autonomous system (when reduction to ODE's is possible) because of covariance. Nowhere is there an explicit dependence on time in the field equations. While this independence of time is a blessing in this case it is, of course, the root of a great deal of trouble in quantum gravity.

Now the evolution equations are: [53]:

$$\begin{aligned}
\dot{\rho} &= -\rho\Theta & (3.5)\\
\dot{\Theta} &= -1/3\Theta^2 - (\sigma_1+\sigma_2)^2 - (\sigma_1^2+\sigma_2^2) - 1/2\,\rho & (3.6)\\
\dot{\sigma}_i &= 2/3\sigma_j(\sigma_1+\sigma_2) - 1/3\,\sigma_i^2 - 2/3\,\Theta\sigma_i - E_i & (3.7)\\
\dot{E}_i &= E_i(\sigma_i-\sigma_j) - E_j(\sigma_i - 2\sigma_j) - \Theta E_i - 1/2\,\rho\sigma_i & (3.8)\\
& & (3.9)
\end{aligned}$$

where we have used a condensed notation with $i, j = 1, 2$ but with $i \neq j$ in all cases. The first equation is just the matter conservation equation for dust, while the second is the Raychaudhuri equation, controlling the expansion rate.

Now in FLRW and Bianchi spacetimes, all physical quantities diverge at singularities so that study of endpoints of collapse involve looking at infinity in the phase space. This turns out to be true in silent universes too, so that more useful variables are those that factor out the expansion [1], thereby bringing the singularity to finite values of the new variables. Further, one should note that the equations (3.9) simplify greatly whenever $\sigma_1 = \pm\sigma_2$ and $E_1 = \pm E_2$. Thus if we define new variables:

$$\Omega = 3\rho\Theta^{-2}\ , \quad \Sigma_\pm = \frac{1}{2}(\sigma_1 \pm \sigma_2)\Theta^{-1}\ , \quad \epsilon_\pm = \frac{1}{2}(E_1 \pm E_2)\Theta^{-2}$$

One can now study the case $\Theta \to \infty$ much more easily, with the values of $\Omega, \Sigma_\pm$ and $\epsilon_\pm$ finite at the singularities. In fact, if we change time variable to:

$$\tau = \pm \int \Theta dt = \pm 3 \ln \ell \tag{3.10}$$

then a further significant simplification occurs: namely the phase space decouples in one direction in the sense that the stationary points of the phase space are determined solely by $\Omega, \Sigma_\pm, \epsilon_\pm$, which no longer depend on the expansion, Θ. The phase space is reduced effectively to 5-dimensional with only the Raychaudhuri equation dependent on the expansion, Θ. The exact equations are [53] (for $\Theta < 0$):

$$\Theta' = \Theta\left(1/3 + 6\Sigma_+^2 + 2\Sigma_-^2 + 1/6\ \Omega\right) \tag{3.11}$$

$$\Omega' = -1/3\Omega\left(36\Sigma_+^2 - 1 + 12\Sigma_-^2 + \Omega\right) \tag{3.12}$$

$$\Sigma_+' = \Sigma_+\left(1/3 - \Sigma_+(6\Sigma_+ + 1) - 2\Sigma_-^2 - 1/6\ \Omega\right) + 1/3\ \Sigma_-^2 + \epsilon_+ \tag{3.13}$$

$$\Sigma_-' = \Sigma_-\left(1/3 - \Sigma_-(6\Sigma_+ - 1) - 2\Sigma_-^2 - 1/6\ \Omega\right) + \epsilon_- \tag{3.14}$$

$$\epsilon_+' = \epsilon_+\left(1/3 - 3\Sigma_+(4\Sigma_+ - 1) - 4\Sigma_-^2 - 1/3\ \Omega\right) - \Sigma_-\epsilon_- + 1/6\ \Sigma_+\Omega \tag{3.15}$$

$$\epsilon_-' = \epsilon_-\left(1/3 - 3\Sigma_+(4\Sigma_+ + 1) - 4\Sigma_-^2 - 1/3\ \Omega\right) - 3\Sigma_-\epsilon_+ + 1/6\ \Sigma_-\Omega \tag{3.16}$$

In these equations, changing from studies of collapse to studies of expansion is only seen in the change in the sign of the Raychaudhuri equation.

3.3.2 Endpoints of gravitational collapse

Because the continuity equation, the Raychaudhuri equation and the Bianchi identities (which give equations for $\dot{E}_{ab}, \dot{H}_{ab}$) are all autonomous (they do not depend explicitly on time), when the spatial gradients are eliminated from these equations the resulting coupled set of ODE's will also be autonomous. This means that a dynamical systems approach can be used very effectively to study global properties of the system such as stability and attractors. In particular we may ask "what is the generic endpoint of collapse in the irrotational silent universe ?"

We can answer this question by looking for the fixed or stationary points of the system of equations given by (3.16). This has been done in [53].

[1] Alternatively one could choose a less physical means of compactifying the phase space using coordinate transformations based on unbounded functions, such as $\tan^{-1} x$ for example.

The Szekeres models

These correspond to the special case where $\Sigma_- = \epsilon_- = 0$, so that the phase space is effectively 3 dimensional. The vanishing of the above two variables means that two of the eigenvalues each of the shear and electric part of the Weyl tensor must be equal. Even so the Szekeres models are generalisations of the Kantowski-Sachs and Lemaître-Tolman-Bondi solutions.

It turns out that for initial conditions with $\Theta < 0$, the generic attracting subspace is the $\Theta = 0$ plane with six physical fixed points. Conversely, if we are interested in expanding models ($\Theta > 0$), we need just reverse the arrows on the trajectories. The stationary points correspond respectively to the flat FLRW model, a conformally flat void locally equivalent to a Milne universe, a Kasner solution with pancake singularity, another Kasner model with spindle singularity and a solution which is the limit of a subclass of Szekeres models. Finally there is a vacuum solution which the authors could not identify [53].

The eigenvalues of the Jacobian of the flow gives the stability of the above solutions: the flat FLRW, vacuum and Szekeres subclass are all unstable (saddle points). The Milne attractor is completely stable and is the final fate of ever-expanding voids. The two Kasner solutions are attractors during collapse for pancake and spindle configurations respectively.

3.3.3 Triaxial dynamics

Having discussed the special case of Szekeres models in the previous subsection, we would like to give a brief overview of the general case representing triaxial dynamics governed by eq.s (3.16).

As in the Szekeres models, the collapse proceeds towards the plane $\Theta = 0$. Because the origin is a stationary point, all the critical points of the Szekeres models will be critical points of the general dynamics. However there are additional critical points lying on Lissajous-like closed curves (which do not self-intersect), representing relativistic triaxial Kasner solutions [53] - one pancake singularity (which is the same as the Szekeres pancake) and the rest spindle singularities. An interesting note is that none of the attractors correspond to the weakly-chaotic mixmaster-Bianchi IX models which exhibit oscillatory spindle singularities along different axes [59]. In particular this partially disproves the conjecture by Belinskii, Khalatnikov and Lifshitz that Mixmaster dynamics are generic at singularities, as suggested by several authors [60], at least for irrotational silent universes. That this will remain true in the vortical case is strongly suggested by the fact that a rotational degree of freedom removes the Mixmaster behaviour from the singularity in Bianchi IX [58].

Finally, although the basins of collapse are more complicated it remains true that ever-expanding volume elements approach the spherical void (Milne) solution.

3.3.4 Comparison with N-body simulations

Here we will compare the silent universe results with those obtained from dissipationless (i.e. dust) N-body simulations which ignore vorticity and gravitational waves. However, they are only followed until roughly the present epoch so one cannot strictly compare the results since in principle the N-body results could transform to a different shape exhibiting a type-of Mixmaster behaviour. Further, in the relativistic case one is studying the evolution of the metric while in the Newtonian case one looks at evolution of the fluid density. Nevertheless, it is a comforting result if the two agree. The generic nature of the spindle singularity is supported by many numerical studies of N-body collapse (e.g. [83]). Two interesting features have appeared: (1) it is the electric part of the Weyl tensor which is the dominant characteriser of structures: the spindle structure is due to a large E_{ab} eigenvalue along the axis of the spindle [83, 84]. This will be relevant to our later discussion on the effect of vorticity. (2) Pancake structures, although less likely, may form before spindle ones.

Given this, the recent work on "Quiet" universes is very interesting [86]. There it was claimed that the spindle configuration is unstable to H_{ab} perturbations, while the pancake configuration is not. If they are correct, then the E_{ab}- dominated spindle shape may be just a passing phase which is later transformed into a different state. In this case, one may ask why it is that numerical simulations have not discovered this ?

3.4 Petrov classifications of Irrotational silent models

Before we embark on a deeper study of silent universes, it will pay us to give a brief overview of the Petrov classifications of the Weyl tensor. This is a rather sophisticated technique in General Relativity and we will not delve into it too deeply, as it is best formulated in terms of spinors, their connections and curvature, and in particular conformal spinors.

Now one can derive a spinor Ψ_{ABCD} from the Weyl tensor which can be decomposed (uniquely up to rescaling) into four spinors, called the principal spinors of the Weyl tensor. Since every spinor defines a real, null vector, the principal spinors, η^A, of the Weyl tensor define four real null vectors known as the principal null directions, k_i, of the Weyl tensor. The principal spinors then satisfy the equation:

$$\Psi_{ABCD}\eta^A\eta^B\eta^C\eta^D = 0 \tag{3.17}$$

This can be formulated equivalently as:

$$k^i k^k k_{[p}C_{i]jk[l}k_{q]} = 0 \tag{3.18}$$

At this stage we are able to look at special cases of the principal spinors η^A. In particular, if two or more of them are proportional to each other then the Weyl tensor is called *algebraically special.* They have been classified according to the possible special cases and the results are shown in table 3.1.

The type N fields are radiative, describing gravitational waves. The type O fields are conformally flat and are all known, while the type D fields are purely "Coulombian" such as the static spherically symmetric solutions satisfying Birkhoff's theorem. Type I fields are the most general and can be thought of as a superposition of purely Coulombian and transverse fields in general [57]. Not all type I fields are inhomogeneous however. The OSH

Bianchi I, for example, is both type I and homogeneous [80]. Bruni *et al* [53] attribute the genericity of the spindle collapse to the type D (Coulomb) field resulting from the initial conditions and constraints on E_{ab} that has a nonlocal effect causing the pancake solutions to be unstable.

3.4.1 A brief discussion of competing effects

The assumptions of the irrotational universe are very likely fulfilled on superhorizon scales. However, because we are dealing with nonlinear perturbations, each mode of the corresponding generalised Fourier problem does not decouple (as it does in the linear case) and all modes will be coupled to all others. Thus if the approximations of the silent universes are violated on small scales, we can expect that the mode coupling may transfer this information to the whole spectrum of wavelengths. Still, one expects from studies of other coupled-mode systems, notably magneto-hydrodynamics, that the mode-coupling will preferentially cascade power to smaller scales. Thus we expect that when we consider the universe on superhorizon scales, the silent universe approximations will probably be fulfilled to high accuracy.

This generalised Fourier approach to the mode couplings in spatial sections is 'orthogonal' in a certain sense to the silent universe formalism which is concerned only with the evolution of single worldlines independent of the rest of the spatial sections except on the initial spacelike hypersurface on which initial conditions were placed (and where the constraint equations must hold). In fact this difference describes geometrically the competition between the effects of spatial gradients which cause mode-coupling via density gradients, sound and gravitational waves, and silent evolution which only cares about initial conditions. On subhorizon scales, even if $H_{ab} = 0$ initially, weak mode-couplings at second order due essentially to pressure gradients, destroy this equality [56]. The further evolution of the universe is a complex, scale-dependent competition between the silent, "initial conditions", and the mode coupling due to spatial gradients.

In the two limits of small and large scales the situation appears fairly well known - silent universes on the one hand and Newtonian gravity coupled to the Navier-Stokes and plasma equations on the other. It is the intervening stage that is the problem. Partly because we don't know where the transition occurs.

3.5 Generation and evolution of vorticity

Using a Lagrangian formalism, Buchert (1992) [69] showed that mass conservation allows the exact integration of the evolution equations to give:

$$\omega = \omega_0 \cdot \nabla \mathbf{f} (det\mathbf{F})^{-1} \tag{3.19}$$

where $\mathbf{f}$ is the flow vector and $\mathbf{F} = \nabla \mathbf{f}$ is the deformation gradient of the flow. Geometrically this is significant because it shows that the vorticity shares caustics (where shell crossings occur) with the density, when $det\mathbf{F} = 0$. This is a generalisation of the famous Zel'dovich solution, except here both the density and the vorticity diverge formally, and at the same spacetime events, so that regions of high density are also expected to be regions of high vorticity.

This geometric demonstration of the possible importance of vorticity makes it important to know the scale on which vorticity is significant, particularly for testing the validity of the POTENT approximation[2], which explicitly requires an irrotational velocity field. The Fourier space problem where the answer lies is complex because of the nonlinearity involved. This implies that different modes will not decouple as in the linear case so that cross-correlations between modes and between dynamical quantities will act as sources and are expected to lead to cascades as in classical turbulence. This makes it too difficult to study in the present work [3], and we simply state the problem, the answer to which will govern the validity of a large number of cosmological studies.

Initial conditions provide the other half of the key to this problem since the vorticity spectrum might have been attenuated so greatly initially that even nonlinear growth and mode couplings could not have transferred it by the present epoch to scales greater than tightly bound galaxy clusters.

On the other hand, vorticity may also be generated, even if it were absent initially, by irreversible processes [75]. In complete generality we have [8]:

$$h^a_b(\dot{\ell^2 \omega^b}) = \sigma^a_c(\ell^2\omega^c) + \frac{\ell^2}{2}\eta^{abcd}u_b\dot{u}_{c;d} \tag{3.20}$$

where h^a_b is the projection tensor, ℓ is the "scale" factor, η^{abcd} the antisymmetric alternating tensor and u_a the four velocity. The last term in eq. (3.20) can be non-zero either due to spatial variations in the pressure gradient, or to the shear or vorticity coupling to a non-zero flux, as exists in realistic multi-fluid models [35]. If the shear is nonlinear, then the growing mode of the vorticity is ensured, since ℓ will be a locally decreasing function in this case. If the last term is zero and the shear is linear however, then the right hand side can be neglected, and we recover the *pressure-free*, decaying mode solution, $\omega \propto \ell^{-2}$, so often used as a justification for neglecting vorticity.

When pressure is included in the linear fluctuations, we see that vorticity may be *created or destroyed* through irreversible processes: [39]:

$$\dot{\omega_{ab}} = -\frac{2}{3}\Theta\omega_{ab} + \left[c_s^2\Theta - (\frac{\partial p}{\partial s})_\rho \frac{\dot{s}}{\rho + p}\right]\omega_{ab} - ^{(3)}\nabla_{[b}\Pi_{a]} \tag{3.21}$$

where ρ is the energy density, $\Pi_a = (\rho + p)^{-1(3)}\nabla^b\pi_{ab}$, π_{ab} is the anisotropic stress, c_s^2 is the speed of sound, s is the entropy and $[ab]$ denotes anti-symmetrisation on those indices. Again we have not included a source term due to fluxes from tilted frames or multiple (i.e. baryon-dark matter) fluids. More physically, vorticity can be generated if ${}^{(3)}\nabla\rho\wedge{}^{(3)}\nabla p \neq 0$, if there are tangential stresses (such as due to a magnetic field, which is another axial vector field with very similar behaviour), or oblique shock waves in the flow. Further insight is gained, in the absence of shear and pressure, from the vorticity evolution equation (3.21) which is dominated by the local value of Θ, the expansion. If the expansion is negative, i.e.

[2]The POTENT scheme is a method for reconstructing the gravitational potential Φ from the peculiar velocity field, since at linear theory with an irrotational velocity field, the two are related by a simple integral.

[3]The weakly nonlinear stages of the irrotational Fourier space problem have been studied recently [79].

local collapse is occurring, the vorticity grows. Now since at linear level, $\Theta \propto -\nabla \cdot \mathbf{v} \propto \delta$; regions of positive density contrast are roughly associated with growing vorticity unless shear or dissipation are dominant. Thus even if our universe had extremely special initial conditions with purely adiabatic fluctuations and with the vorticity *exactly* zero everywhere, this would not have remained the case and we need to return to the Fourier space problem to discover on what scales vorticity is important.

3.6 Vorticity in the early universe - constraints and generation

The main reason, apart from mathematical completeness, for considering vortical silent universes stems from the fact that vorticity may be important in the dynamics of certain scales in the universe, and a basic mistrust of the general "rule-of-thumb" that one can ignore vorticity. The main question that we must answer then, is how could vorticity have been produced in the early universe ? We have given some general ideas in the previous section on vorticity generation but here we will concentrate on specific models.

It is well known that turbulence cannot have seeded galactic structures since it would have caused anisotropies in the CMB above the observed $\Delta T/T \simeq 10^{-5}$ and would certainly have violated the strict COBE FIRAS constraints on spectral deviations from blackbody (see chapter 4). Conversely however, nothing stops part of the CMB anisotropy from being due to a spectrum of vector perturbations (see chapter 4 for a discussion of how vorticity creates anisotropy in the CMB). The question arises, how would such a spectrum be created ? Since inflation, topological defects and the Primordial Baryon Isocurvature (PBI) models are the main possible ways of seeding structure known today we will ask the question of these theories.

Now because the vorticity of scalar fields is identically zero, it is generally believed that inflation removes any trace of vorticity initially present. This is true bar two caveats: firstly we expect the inflaton field to be coupled to other fields which most likely support vorticity. Thus as discussed by Grischuk [40], vortical perturbations in these fields could be parametrically amplified by the expansion of the universe. Secondly, at the end of inflation, reheating must take place where the great entropy of the universe is produced as the inflaton field decays to other fields, again which will generally support vorticity. From eq. (3.21) we see that, depending on the sign of the third term, reheating will either cause a drastic reduction, or increase of vorticity. We know that $\dot{s} \gg 1$ during reheating, so that we are left with the term $(\partial p/\partial s)_\rho$ - the derivative of the pressure with respect to the entropy density at constant energy density. If this is negative the vorticity will have a rapid growing mode and will violate eq. 3.21, which only holds for linear ω, within a few oscillations of the inflaton field at the bottom of its potential. If density perturbations were still small then this vorticity would then decay as usual, but because of the huge change in entropy ($\simeq 10^{40}$), would very likely cause a visible effect in the CMB. Conversely, if $(\partial p/\partial a)_\rho > 0$, $\dot{\omega} < 0$.

What about primordial baryon isocurvature (PBI) models ? These are open, baryon dominated models in which the primordial perturbations were isocurvature entropy perturbations rather than adiabatic perturbations. The weakness of this theory is that the origin of the perturbations is not specified. However, assuming that the entropy perturbations were produced somewhere near decoupling, the same argument as in the inflationary case, but perhaps in a milder form, holds.

What about topological defects ? There are two candidates here: cosmic strings and global textures. Both seed structures via isocurvature entropy perturbations and hence may, depending on the models (as in the cases above), give rise to growing vortical modes. In the case of cosmic strings there is another interesting possibility. Since strings have extremely nonlinear densities and move relativistically, they are supersonic and would produce large-amplitude shocks in the surrounding media. This turbulence would induce vorticity and hence seed magnetic fields [61].

Of the above theories, only cosmic strings can reasonably produce vorticity after recombination, and given the stringent constraints from the CMB, the above discussions should perhaps be taken as constraining those theories rather than offering a way to produce a vorticity component that would still be viable today. As for cosmic strings, it should be possible to place strong limits on the number density of strings per Hubble volume from FIRAS. This is because turbulence and acoustic waves transfer energy to the CMB and hence cause spectral distortions, which can be used to constrain the spectral index of the power spectrum, even in standard inflationary models [74].

Of course we have neglected standard hydrodynamical effects which will definitely generate vorticity on smaller scales, and as discussed earlier, once local collapse occurs, the vorticity gains a growing mode.

3.7 Vortical silent universes (VSU) ?

Returning to the subject of the silent universes, it is instructive to ask the question: under what conditions can the silent nature of the flow be retained ? Here we conjecture that the irrotational assumption is useful, not for ensuring the silentness of the flow, but rather for ensuring the existence of a simultaneous diagonalisation of the shear, electric part of the Weyl tensor, E_{ab}, and metric, and hence a reduction of the full field equations to just six ordinary differential equations (ODE's) [53]. However, before discussing this conjecture, let us examine the evidence *against* it.

Due to Poincaré invariance, there are strong similarities between gravity and electromagnetism (EM). Let us examine this analogy with electromagnetism (EM). If there were magnetic monopoles then these would be a source of the magnetic field just as electric charges are a source of the electric field. Hence if one wanted to eliminate the magnetic field, one would also have to eliminate magnetic monopoles. Now gravity is a non-Abelian theory so "monopoles" are possible. What are the gravitational monopoles ? Looking at the *linearised* div H constraint equation (2.19) we see that div $H_{ab} \neq 0$ when $\omega_a \neq 0$ (if $\mu \neq 0$). So that vortices are the monopoles of non-vacuum, linearised gravity. This is not strange, since it is also the case in other gauge theories. Indeed in chapter 2 we imposed the constraint that $\omega_a = 0$, i.e. that there were no vector perturbations. But here we want that there be no gravitational waves (so that there is no spatial communication). Now if the analogy with EM carries through, allowing vorticity in silent universes would create a "Coulombian" magnetic Weyl tensor, and hence vorticity would act as a source of H_{ab}. The question would then be, is it possible to have $H_{ab} \neq 0$ and still not have gravitational waves ?

That the analogy with EM carries through to gravity appears to hold very well. The recent work of Bonnor [81] on the invariants $E_{ab}E^{ab} - H_{ab}H^{ab}$ and $E_{ab}H^{ab}$ show no flaws in the analogies between GR and EM. Further, another work by Bonnor [82] demonstrates that at least in the special case of the van Stockholm solution, vorticity does act as a source for the magnetic part of the Weyl tensor. Hence, while in terms of the physicallity of the models, adding vorticity improves the situation, it may worsen the integrability problems of silent models.

3.7.1 Simultaneous diagonalisations

In the irrotational case we used the "div H" equation (eq. 3.1) to prove that the eigenframes of the shear and E_{ab} coincided. Unfortunately we can see that this is no longer true in general when vorticity is included as the "div H" equation now becomes (assuming no spatial gradients) [8]:

$$\eta^{tbpq} u_b \sigma^d_p E_{qd} = 3E^t_s \omega^s + \rho\omega^t \tag{3.22}$$

so that in general, the commutator between the shear and E_{ab} matrices is not zero. Hence the symmetry between the eigenframes is broken and E_{ab} and σ_{ab} cannot be simultaneously diagonalised in the general vortical case. Alternatively we may formulate it as:

<u>**Lemma**</u>

For purely Weyl-electric dust ($p = H_{ab} = 0$), the shear σ_{ab} and electric part of the Weyl tensor, E_{ab} have the same eigentetrad and hence are simultaneously diagonalisable if $\omega_{ab} = 0$.

Conversely, if $\omega_{ab} \neq 0$ then a simultaneous diagonalisation is only possible when $3E^b_a\omega^a + \rho\omega^b = 0$. i.e. when the vorticity vector is an eigenvector of E_{ab} with continuous eigenvalue $-\frac{1}{3}\rho$.

<u>*Proof*</u>

The proof follows immediately from eq. (3.22) on setting the RHS $= 0 = p$.
The special case with non-zero vorticity mentioned above corresponds geometrically to the case where the instantaneous axis of rotation of the fluid is aligned with one of the principal directions of E_{ab}, and hence of the shear. This means that the vorticity rotates the two non-aligned principle axes of σ_{ab} and E_{ab} in a plane orthogonal to ω_a. In this case our extended set of coupled ODE's becomes 9 dimensional, with the addition of three equations for the off-diagonal terms of ω_{ab}. There are no diagonal terms because $\omega_{ab} = -\omega_{ba}$.

However, for general ω_{ab} it is not possible to achieve a simultaneous diagonalisation of σ_{ab} and E_{ab}. This means that we can diagonalise one of σ_{ab} or E_{ab} by suitable choice of tetrad, but not both. This in turn implies that we need three more equations for the extra off-diagonal components of either σ_{ab} or E_{ab}, bringing the total of equations to twelve.

In the general case one is left with a choice of tetrad which coincides with the eigentetrad of the shear, electric tidal field, or such that one of the spatial tetrad vectors is parallel to the vorticity vector. More specifically one can always choose the tetrad such that $\omega = \omega_1$, $\omega_2 = \omega_3 = 0$.

3.7.2 The evolution equations for VSU's

As before we have matter conservation:

$$\dot{\rho} = -\rho\Theta \tag{3.23}$$

and the modified Raychaudhuri equation:

$$\dot{\Theta} = -\frac{1}{3}\Theta^2 - 2\sigma^2 + 2\omega^2 - \frac{1}{2}(\rho + 3p) \tag{3.24}$$

These are easily converted to tetrad form since they are scalar equations. The rest of the relevant equations are simplified by the fact that, even if not simultaneously diagonalisable, the projection tensors in the time derivative terms can be dropped due to the geodesic (pressure free) flow (G.F.R. Ellis, *Pvt. Comm.*). We therefore have the vorticity evolution equation:

$$(\omega^a\dot{)} = \sigma^a{}_b\omega^b - \frac{2}{3}\Theta\omega^a \tag{3.25}$$

In our case the acceleration is zero and the equation is in fact simply a conservation equation for $(\ell^2\omega^b)$, where ℓ is the characteristic length (see equation 3.4). In addition there is the shear propagation equation with two extra terms relative to the irrotational case, one proportional to $h_{ab}\omega^2$ and one to $-\omega_a\omega_b$:

$$(\sigma_{ab}\dot{)} = -\omega_a\omega_b - \sigma_{af}\sigma^f_b - \frac{2}{3}\Theta\sigma_{ab} + \frac{h_{ab}}{3}(\omega^2 + 2\sigma^2) - E_{ab} \tag{3.26}$$

where $\omega^2 = \omega^a\omega_a$. The Bianchi identities, $R_{ab[cd;e]} = 0$, give us the Maxwell-like equations [8]:

$$\begin{aligned} \dot{E}^{ab} &= -J^{ab} - h^{ab}\sigma^{cd}E_{cd} + \\ &- \Theta E^{ab} + 3E_s{}^{(a}\sigma^{b)s} + E_s{}^{(a}\omega^{b)s} - \frac{1}{2}(\rho)\sigma^{tm}, \end{aligned} \tag{3.27}$$

for the time variation of E_{ab}. This is the set of equations (with $J^{ab} = 0$) which describe the dynamics of VSU's. The last part of our job is to write the projection tensor h_{ab} in terms of our other kinematic variables.

Note that in general we will have the evolution equation for H_{ab} which in our case becomes a constraint:

$$\begin{aligned} h^m{}_a h^t{}_c \dot{H}^{ac} - I^{mt} + 2E_a{}^{(t}\eta^{m)bpq}u_b\dot{u}_p + h^{mt}\sigma^{ab}H_{ab} + \\ +\Theta H^{mt} - 3H_s{}^{(m}\sigma^{t)s} - H_s{}^{(m}\omega^{t)s} = 0 \end{aligned} \tag{3.28}$$

where we have followed the modern trend [8, 49] of illucidating the importance of the curl-like terms responsible for wave-like motion of gravitational waves by writing:

$$J^{mt} = h_a{}^{(m}\eta^{t)rsd}u_r H^a{}_{s;d} = \text{“}Curl\ H\text{”} \tag{3.29}$$

$$I^{mt} = h_a{}^{(m}\eta^{t)rsd}u_r E^a{}_{s;d} = \text{“}Curl\ E\text{”} \quad (3.30)$$

which by the definition of silent universes are zero. These then are the equations (without the $\dot{H}_{ab}$ eq.) that must be converted to tetrad form if we wish to study the system from a dynamical systems point of view. If we examine these equations we see that the only technical difficulty is provided by the presence of the projection tensor, $h_{ab} = g_{ab} + u_a u_b$. Now the theorem quoted earlier [55] for the irrotational case guaranteed that the metric was generically (except in two special cases), diagonalisable with space-like components ℓ_α. In the irrotational case this is not expected to hold.

As a final note, we cannot use the Gauss-Codazzi equations since these rely on a $3+1$ decomposition which is strictly only possible when the vorticity vanishes, although it is possible to define them at first order if the vorticity is linear.

3.7.3 The integrability conditions

For any solution to the propagation equations, one must check that the constraint equations are also satisfied, i.e. a full solution to the ten field equations. In particular one must check that the constraint equations are preserved under time evolution. In this regard, Newtonian solutions are bad rôle models. A well-known example is provided by homogeneous rotation, which appears to allow the avoidance of the initial singularity, but which does not satisfy the constraint equations [8].

Now, it is widely believed that General Relativity is in fact an integrable system, among other reasons because it has a formulation in terms of twistors [62, 63], as do 4-dimensional self-dual Yang-Mills theories. However, proofs of these beliefs are still pending and in the mean time one must proceed from case to case. In vacuum, $T_{ab} = \omega_{ab} = 0$, the Bianchi identities imply that the constraints are preserved under the flow. The corresponding case for the irrotational silent universes had been thought to be proven in [54]. However, the converse actually appears to be likely in general. There is a conjecture that there do not exist any truly inhomogeneous Petrov type I models [80]. This is particularly interesting since, if true, it implies a linearisation instability in the field equations since silent linear perturbations of FLRW models are known to satisfy the constraints identically. Hence there may exist linearised solutions which have no correspondence in the full non-linear theory.

The further question under debate here is whether or not the silent flow with vorticity observes the constraint equations. The constraints come from the Ricci identities ($u^a_{;bc} - u^a_{;cb} = R^a_{dbc}u^d$). One obtains conservation equations for the shear, vorticity and Raychaudhuri equations together with the three constraints

(a) The $(0,\nu)$ equations:

$$h^{ab}(\omega_{b;c}^{\ c} - \sigma_{b;c}^{\ c} - \frac{2}{3}\Theta_{;b}) = 0 \quad (3.31)$$

(b) the vorticity divergence constraint:

$$\begin{aligned} h_a{}^b\omega^a{}_{;b} &= \omega^a\dot{u}_b = 0\,, && (3.32)\\ \Longrightarrow\quad \nabla\cdot\omega &= 0 && (3.33)\end{aligned}$$

$$\Longrightarrow \quad \omega \quad = \quad \nabla \times \mathbf{A} \tag{3.34}$$

$$\tag{3.35}$$

Thus in the case where there is no acceleration (pressure-free, geodesic motion), we have a vorticity vector potential **A**.

(c) The H_{ab} constraint:

$$H_{ad} = 0 = h^t_a h^s_d (\omega_{(t}{}^{b;c} + \sigma_{(t}{}^{b;c})\eta_{s)fbc} u^f \tag{3.36}$$

where we have consistently put $p = 0 = \dot{u}_a = q_a$, where q_a is the flux term in the stress tensor of a non-perfect or multi-fluid.

In addition there are the "Maxwell" constraints for "div E" and "div H":

$$h^t{}_a E^{as}{}_{;d} h^d{}_s - \eta^{tbpq} u_b \sigma^d{}_p H_{qd} + 3H^t{}_s \omega^s = \frac{1}{3} h^{tb} \rho_{;b} \tag{3.37}$$

and

$$h^t{}_a H^{as}{}_{;d} h^d{}_s + \eta^{tbpq} u_b \sigma^d{}_p E_{qd} - 3E^t{}_s \omega^s = \rho \omega^t \,, \tag{3.38}$$

where the first is trivially satisfied in silent universes and the second gives us the constraints on simultaneous diagonalisation, as discussed previously.

<u>The $H = 0 = \dot{H}$ Silent constraint</u>

Now this "silent" constraint, because the vorticity only comes into it in a product with H_{ab}, *is exactly the same as in the irrotational case*:

$$0 = h_a{}^{(m}\eta^{t)rad} u_r E^a{}_{s;d} \tag{3.39}$$

Only once the time derivative of this is taken do any differences appear. The two new terms give the constraint [54]:

$$0 = h^{(i}{}_n \eta^{j)klm} u_k \left[(\dot{E}^n{}_l)_{;m} - E^n{}_{l;p} u^p{}_{;m} \right] \tag{3.40}$$

Now of these two terms, the first will introduce an extra term related to

$$(E_s{}^{(m}\omega^{t)s})_{;n}$$

while the second is unknown. Thus the problem with this constraint is reduced to checking that equation (3.40) holds under time propagation.

There are two issues here: that vorticity does not destroy the silent nature of the flow, and secondly, that $p = H_{ab} = 0$ is consistent with $\omega_a \neq 0$ - integrability conditions for the flow. That vorticity doesn't destroy the silent nature is seen from the lack of spatial gradients in the equations under the assumptions $p = H_{ab} = 0$ and physically understandable since communication (through sound or gravitational waves) between world lines is not dependent on vorticity. The integrability conditions, on the other hand are not proven here, and will likely remain so for a good length of time, given the complexity of the problem.

Finally, it seems likely that it is possible to obtain silent universes with $H_{ab} \neq 0$. From chapter 2, it is clear that if we want to eliminate gravitational waves we should rather impose the vanishing of the curl's of E_{ab} and H_{ab}, thus leaving the Coulombic part of the field involved in gravitational collapse. However, this also remains unproven at this time.

3.8 Averaging and the Cauchy Problem

The main problem of introducing vorticity into the cosmological problem, apart from the increased complexity of the field equations, is however, the definition of time surfaces. No longer is it possible to define "time surfaces" everywhere orthogonal to the fluid flow, u^a [8]. However, if initial data are set up when the perturbations are in the linear regime, when the vorticity is linear, as is required from the CMB, this presents no problems, since one is still allowed continuous time surfaces everywhere orthogonal to the fluid flow [8]. This is because the vorticity decouples from the other fluid variables at linear order. However, as soon as the density contrast reaches the quasi-linear and early nonlinear stages, the growing vorticity mode implies that the orthogonality of the fluid flow to the time surfaces is progressively lost.

This is perhaps one of the most interesting problems related to averaging of the field equations: the existence of orthogonal time surfaces on one scale and not on another. It is a crucial problem for studies which use the irrotational velocity assumption for dynamical studies. In particular, the POTENT formalism requires irrotational fluid flow to be valid. Apart from questioning the validity of this assumption on scales of quasi-linear density contrast, it is not obvious that averaging over the small scale nonlinearities yields a metric, which, not only has well defined time-surfaces with the fluid flow orthogonal to them, - but yields a *Newtonian metric* which implies the existence of a preferred time slicing (i.e. the scalar Bardeen metric in the longitudinal gauge).

Let us consider for example, a FLRW model with a nonlinear clump imbedded in it, with the density still finite. The vorticity is assumed zero when averaged over large enough scales, and hence there exists a global time surface everywhere orthogonal to the *averaged* fluid flow. However, as the averaging scale is reduced, vorticity appears near the object and when the vorticity becomes nonlinear, the time surfaces cannot maintain their orthogonality to the flow *and* mesh together to form a single smooth surface. What are the physical consequences of this ? The most important question regards commutativity: if we average over the vorticity and then propagate this solution forward, and then evolve the small scale solution forward (which has the nonlinear vorticity) to the same stage, and then average on the appropriate large scales, will we obtain the same results ? Because the time and fluid flows differ in such a fundamental way, the answer might be expected to be no. Indeed it is an open question as to whether a simple "polarisation" tensor can redress the imbalance due to ignorance of small-scale dynamics. This casts doubt on whether any study of nonlinear structure which ignores vorticity can be extended forward in time consistently.

3.9 Vorticity and the nature of the singularity

Leaving behind the Cauchy problem, we move to the study of the singularity. Since the main aim of casting the equations in the form of an autonomous system is to use the well developed theory of dynamical systems to study attractors - i.e. the endpoints of gravitational collapse. In particular we would like to know whether the attracting set found in the irrotational silent case [53] are stable to the introduction of vorticity. The vortical silent universe (VSU) attractors for ever-expanding models will almost certainly be the same as in the irrotational case - empty Milne solutions - since the vorticity will monotonically decay away.

However, for collapsing models the situation is more complex because (section 3.5) the vorticity starts growing once local collapse occurs. As the singularity is approached the vorticity diverges, essentially because of the vortical conservation law (3.19). At least in first order (Newtonian) Lagrangian theory, the vorticity and density will share the same caustics. It is not known if this is true in the full, relativistic theory. At any rate it seems very plausible that the VSU's will have a different, and certainly more complex, attracting set, one of which may be the Gödel solution (even if measure zero initial conditions lead to it). In particular the most important question is: will the generic end-point of collapse, which in the irrotational case was the *spindle*, Kasner solution, be changed to a *pancake* singularity ? Or will it remain a spindle singularity belonging to a more complex spacetime ? This question is currently under study.

The apparent divergence of the vorticity at singularities should be taken with some caution, however, when studying effective dynamics. Although the density in collapsing models diverges, the analysis of [[53], section 3.3.2] shows that the attractors are generically *vacuum* Kasner models - i.e. the energy density is not felt by the asymptotic dynamics. Perhaps the vorticity is unimportant in the same way ?

If we examine simpler models with vorticity we can get an idea as to the answer to these questions. The rotational Bianchi IX models [58] show that the vorticity enters the Hamiltonian in two ways: it appears in the matter term as the "rotation wall". Since the universe cannot collide with this wall, rotation in this form doesn't affect the qualitative behaviour near the singularity. In this sense it supports our conjecture above about the lack of importance of vorticity at the singularity.

However, rotation enters the Hamiltonian of rotating Bianchi IX's another way - through a centrifugal term which acts on the geometry in a way that matter cannot. Despite this possibility it was found that rotation doesn't in fact change the essential nature of the singularity relative to the non-rotating case. However the Bianchi models have of course homogeneous spatial sections, or in the rotating case, 3-d subspaces, so that no inhomogeneous modes are excited. Now in our inhomogeneous vortical silent universe, the energy density corresponding to the vorticity should be essentially unimportant. However the constraint equations, particularly the $(0,\nu)$ equations (3.31) and the H_{ab} constraint, link the vorticity to the geometry in the way that matter cannot. In particular, we see that the $(0,\nu)$ constraints relate the h^{ab} - projected divergence of the vorticity to the h^{ab} - projected divergence of the shear and covariant derivative of the expansion. This coupling of "spatial" derivatives may be much more important in inhomogeneous models than in the Bianchi models and hence could mean a different result - that vorticity does change the singularity qualitatively. This possibility is suggested by the case of the Kerr model of the rotating black hole. The introduction of rotation deforms the Schwarschild point singularity into a spacelike closed curve singularity which allows for closed timelike curves, which is qualitatively different from the non-rotating case.

In studies of the AVTD Gowdy models (see section 3.9.1), it was found that there is a complex interplay between nonlinear steapening and "freezing" of structures. We know that there is no chaos in the irrotational silent universes, but it should be possible to search for such frozen nonlinear structures on trajectories which allow them - e.g. the Szekeres models.

Now, if vorticity is important then it can induce at least two changes:

(a) the generic singularity reverts to being of *pancake* type.

(b) the generic singularity is locally of oscillating, Mixmaster type (the BKL [64] conjecture).

The problem with accepting (b) is that the existence of the necessary global time-slicing is debated [64, 65]. With vorticity this problem is even more severe, and if this is true in general, then the spatial structure of the singularity becomes extremely complicated as the characteristic bounces occur at different locations at different times. This would be a much more complex version of the singularity in the Tolman universe where the singularity hypersurface occurs at different times depending on radial coordinate.

3.9.1 Asymptotically Velocity Dominated Models

Although not noted before, there is an intimate link between silent universes and another branch of singularity study in General Relativity - the so-called *Asymptotically Velocity Term Dominated* (AVTD) models, which are characterised by the dominance of time derivatives (velocity dominated) over all spatial derivatives as the singularity is approached [64]. Thus effectively all spatial gradients become negligible. Because only the singularity structure is important (i.e. the limiting behaviour), there is no need to have the spatial gradients exactly zero. Hence the AVTD models can have a non-vanishing magnetic part of the Weyl tensor at finite time before the singularity.

Of course the price this exerts is that one must deal with partial differential equations which must be solved numerically [66], and then only for simple solutions such as the Gowdy models. Even then, the numerical results can only be taken as suggestions of AVTD behaviour and can never be real proofs. It is comparisons such as this which bring out the real power of the silent universe formalism for obtaining concrete results. A slight advantage of the numerical approach is that the constraint equations can be checked numerically (at a given accuracy level) at each time step, thus circumventing the analytical integrability issue.

AVTD singularity of the Gowdy and U(1) models

Here we will very briefly outline numerical investigations of the polarised and unpolarised Gowdy models to give a flavour of current research into AVTD singularities [66, 64]. We will not discuss the details of the numerical solution which involve symplectic integrators designed to handle steep gradients and preserve the constraint equations.

The Gowdy model on $T^3 \times R$ is a vacuum model given by the metric:

$$ds^2 = e^{\lambda/2}e^{\tau/2}(-e^{2\tau}d\tau^2 + d\theta^2) + e^{-\tau}[e^P d\sigma^2 + 2e^P Q d\sigma d\delta + (e^P Q^2 + e^{-P})d\delta^2] \tag{3.41}$$

where λ, P, Q are only functions of θ and τ. The angular coordinates are σ, δ which are periodic, and τ is the time coordinate. This metric admits three Killing fields. If P, Q

are assumed to be small, then they can be identified as the amplitudes of the + and × polarisations of gravitational waves propagating on a background spacetime determined by λ. The case $Q = 0$ is called the polarised model, and can be solved exactly. In general, however, the field equations can be split into two nonlinear "wave" equations for P, Q and two constraint equations. The wave-like equations are [66]:

$$P_{,\tau\tau} - e^{-2\tau} P_{,\theta\theta} - e^{2P}(Q_{,\tau}^2 - e^{-2\tau} Q_{,\theta}^2) = 0 \tag{3.42}$$

and

$$Q_{,\tau\tau} - e^{-2\tau} Q_{,\theta\theta} + 2(P_{,\tau} Q_{,\tau} - e^{-2\tau} P_{,\theta} Q_{,\theta}) = 0 \tag{3.43}$$

Note that as $\tau \to \infty$ (the singularity), the spatial gradients are suppressed by $e^{-2\tau}$ factors, leaving equations which plausibly depend only on τ; hence the AVTD conjecture. In the polarised case the solution has been shown to be AVTD analytically [67]. Detailed numerical calculations for the pseudo-polarised case show very nonlinear structures in P and Q (see figure 3.1) [64].

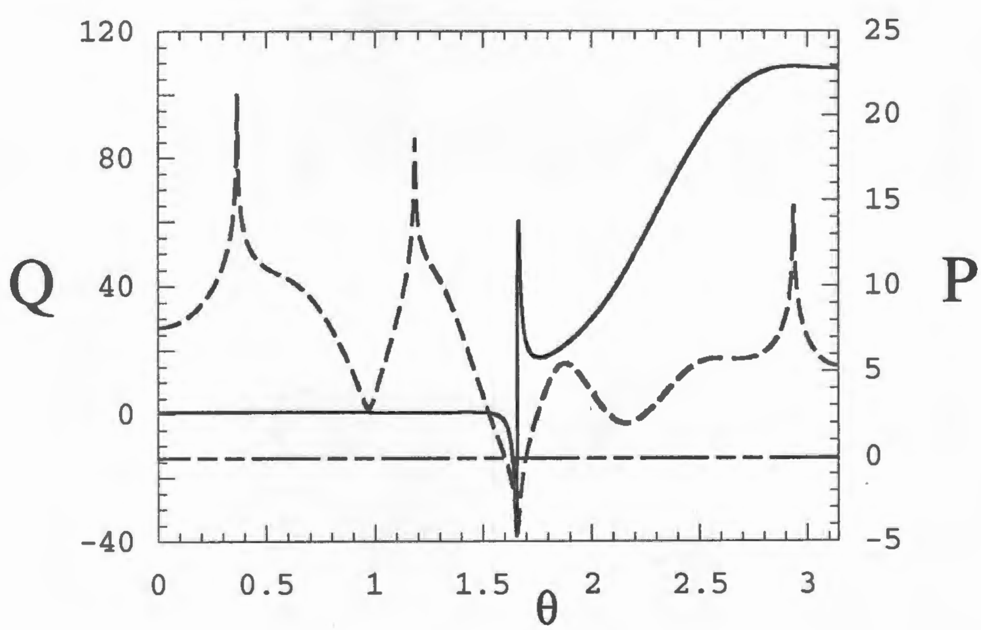

Figure 3.1: P (dashed line) and Q (solid line) vs θ at $\tau = 12.4$ for the standard initial data set with $v_0 = 5$ for $0 \leq \theta \leq \pi$ for a simulation containing 20000 spatial grid points in the interval $[0, 2\pi]$. The peaks in P are essentially the same in that they occur where $Q_{,\theta} \approx 0$ while the apparent discontinuity in Q occurs where $\pi_Q \approx 0$ and $P < 0$ ($P = 0$ is the horizontal line). From [64].

One may extend the numerical analysis to more general cases of interest. In particular it has been shown that any cosmological model on $T^3 \times R$ containing a spacelike $U(1)$ symmetry can be described by only five degrees of freedom. It turns out that the corresponding Hamiltonian formulation is similar to that of the Gowdy models, barring the appearance of a complex term involving all spatial derivatives, characterised by a potential U, which tends to zero as the singularity is approached, in the case of AVTD models.

Numerical results for this more general model are plagued by steep gradients which can only be overcome by spatial averaging which causes deviations from the true dynamics. However, the AVTD behaviour is evident ($U \to 0$) in the following figures (from [64]).

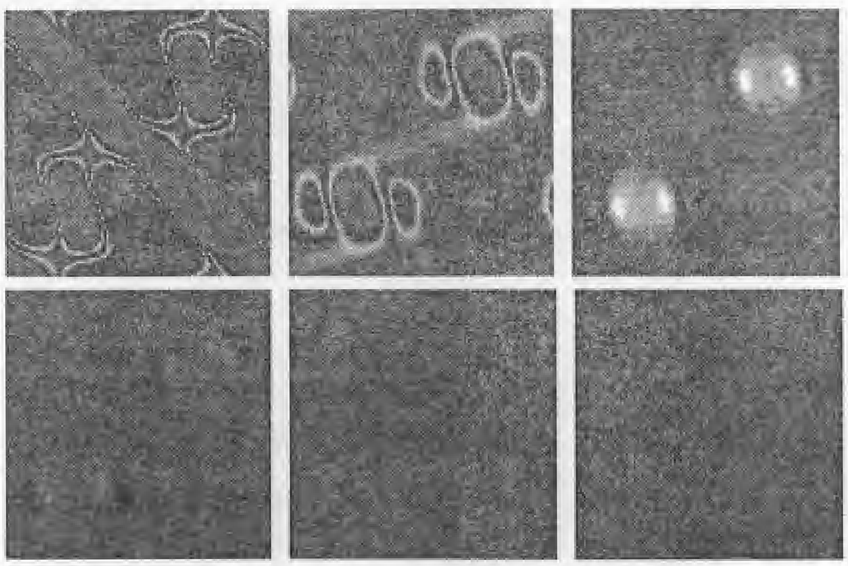

Figure 3.2: Frames of $U(u, v, \tau)$ for the polarized model $x = z = \Lambda = \sin u \sin v$, $p_\Lambda = 12e^\Lambda$, $\omega = r = 0$. Time increases to the right and downward. The final frame corresponds to $U \approx 0$ everywhere [64].

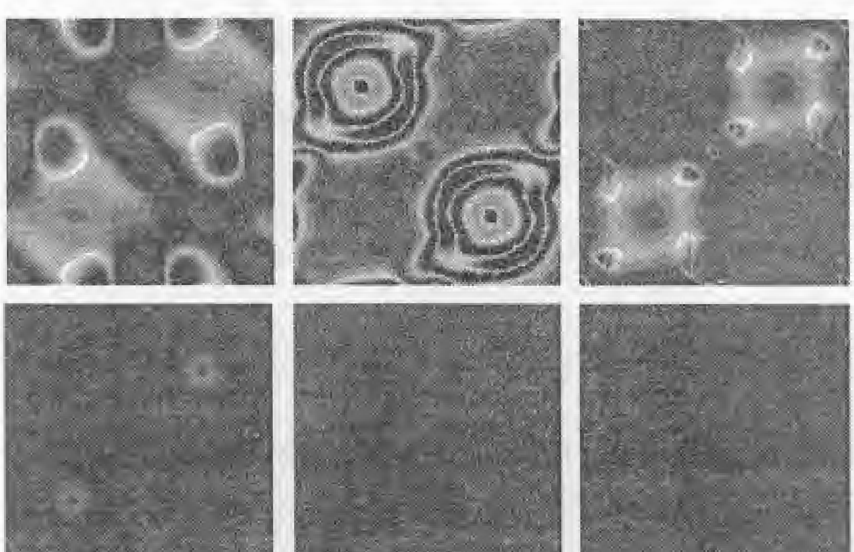

Figure 3.3: Frames of $U(u, v, \tau)$ for the generic model $x = z = \cos u \cos v$, $\Lambda = \sin u \sin v$, $p_\Lambda = 14e^\Lambda$, $p = 10 \cos u \cos v$ with averaging [64].

One notable feature of these models [66] is the development of small scale features due

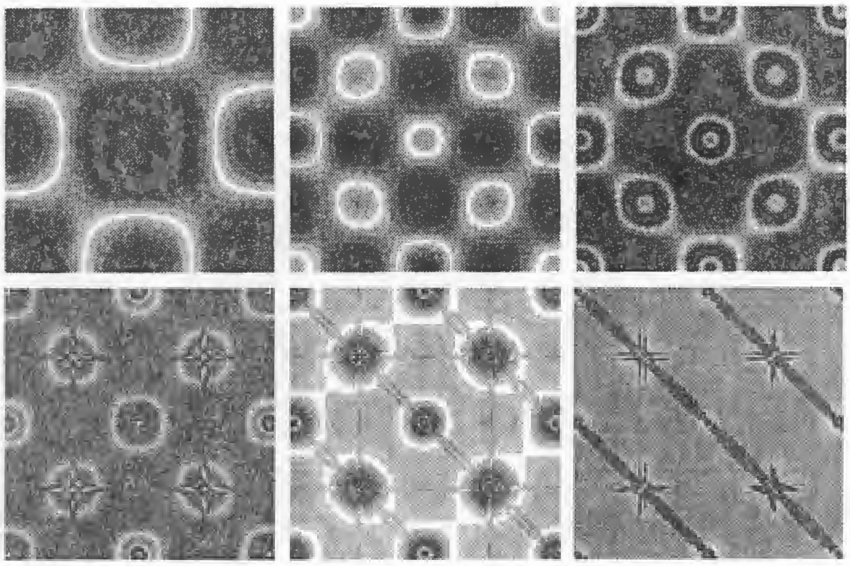

Figure 3.4: Frames of $U(u,v,\tau)$ for generic model $x = z = 0$, $\Lambda = .1\cos u\cos v$, $p_\Lambda = 2.1e^\Lambda$, $r = \cos u\cos v$. The diagonal features in the final frames are numerical artifacts [64].

to nonlinearity which then freeze out as the AVTD limit is approached. This is reminiscent of solitonic solutions of standard nonlinear equations where the steepening due to nonlinearity is balanced by dissipation or dispersion. It would be very interesting to examine eq's (3.42,3.43) for such solitonic solutions.

What is the relationship between AVTD models and silent universes ? Near the singularity, exactly AVTD models can be thought of as a different, spatially homogeneous cosmology at each point in space [66]. This is reminiscent of the silent universe where each worldline evolves as its own, *inhomogeneous*, universe. In this sense the silent models are more general than the exact AVTD solutions. However, the major class of interesting solutions which are AVTD, can only be thought of as classes of homogeneous cosmologies one for each point, exactly at the singularity. In general, particularly away from the singularity, they will have non-zero pressure, magnetic Weyl tensor and spatial gradients, and in this sense are much more general than silent models. Thus we see that the two classes overlap to some extent, with results from one class complementing those from the other.

3.10 The effect on light propagation

How would the existence of vorticity affect light propagation ? The behaviour of an infinitesimal bundles of null geodesics can be formulated through the expansion and (complex) shear of the bundle, with definitions (see chapter 5):

$$\frac{1}{2}\sigma_{\alpha\beta}\sigma^{\alpha\beta} = |\sigma_R|^2 \qquad \theta_R = \frac{1}{2}k^\alpha{}_{;\alpha}$$

where the subscript R denotes the ray-bundle quantity and $\pm|\sigma|$ are the eigenvalues of the shear tensor. These evolve according to the nonlinear optical scalar equations: [142]:

$$\dot{\theta}_R + \theta_R^2 + |\sigma_R|^2 = \frac{1}{2} R_{\alpha\beta} k^\alpha k^\beta \tag{3.44}$$

$$\dot{\sigma}_R + 2\theta_R \, \sigma_R = \frac{1}{2} \, C_{\alpha\beta\gamma\delta} \epsilon^{*\alpha} \epsilon^{*\gamma} k^\beta k^\delta \tag{3.45}$$

where $C_{\alpha\beta\gamma\delta}$ is the Weyl tensor responsible for tidal distortion and $k_\alpha, \epsilon^{*\beta}$ are respectively the wave vector, and a complex null vector in the plane orthogonal to the ray bundle.

The angular diameter distance, [142] (reciprocal to the luminosity distance), r relates the physical size of an object, to the angle, that it subtends in the sky, and is given by:

$$\frac{1}{r}\frac{dr}{d\nu} = \theta_R \tag{3.46}$$

where ν is an affine parameter. This defines the field/fluid approach to gravitational lensing as opposed to the lens plane, point- mass approach, and is perhaps better suited to the study of realistic lensing in multi-fluid baryon-dark matter models beginning to be studied. The impact of vorticity on the ray bundle is not well understood, other than the distortion of critical and caustic curves investigated in simple cases [85]. Under the approximations of the VSU's, the Weyl tensor is given by:

$$C_{abcd} = (\eta_{abpq}\eta_{cdrs} - g_{abpq}g_{cdrs}) u^p u^r E^{qs} \tag{3.47}$$

where

$$g_{abcd} \equiv g_{ac}g_{bd} - g_{ad}g_{bc} \, . \tag{3.48}$$

Taking the time derivative of eq. (3.45) brings into the ray-shear evolution the equation for E_{ab}(again using $\dot{u}_a = H_{ab} = 0$) [8]:

$$\dot{E}^{mt} + \Theta E^{mt} + h^{mt}(\sigma^{ab}E_{ab}) - 3E_s^{(m}\sigma^{t)s} - E_s^{(m}\omega^{t)s} + \frac{\kappa}{2}\rho\sigma^{tm} = 0 \tag{3.49}$$

where a single fluid with comoving observer is assumed so that there is no flux and we assume that there is no anisotropic stress. Naively we might expect that, since the shear and vorticity couple to E_{ab} with the same sign, they have the same effect on the ray bundle. However, the other terms involving the shear - particularly the last one which couples to the density, and is hence expected to dominate most often in the quasi-linear regime - occur with the opposite sign to the vorticity term. Hence we may expect that vorticity will generally act to reduce the convergence of the ray bundle, in contrast to the effects of shear which enhance caustic formation [142]. When one also includes the fact that vorticity slows up the process of gravitational collapse and hence reduces the mass density available to lensing, we see that the net effect of vorticity is probably to *increase the effective* critical surface mass density of a mass distribution [142]. In other words it reduces the possibilities of formation of multiple images of background sources. This *may* be important in using multiple image statistics to constrain cosmogonic models [78]. These involves tests such as : (1) finding the most likely value for the angle of image splitting for sources of different redshifts, (2) the most likely value of the lens redshift (3) the largest expected image splitting angle.

Vorticity slows down collapse, and hence results in less nonlinearity and hence will skew the first of these to smaller angles for almost any model, thus allowing better fits with observations [78]. For the same reason it is expected to bring the lenses to a lower redshift, possibly contrary to observations [78].

3.11 Density Waves in Silent Universes

Density waves are a fundamental part of spiral arm formation in galaxies, but their application to cosmology has been much less investigated. Even in the linear case, density waves - movement of local density extrema from one worldline to others - occur in flat geometries, [77] unless $\frac{dp}{d\rho} = 0$, implying that there are no entropy perturbations, or sound waves. In the weakly nonlinear and fully nonlinear regimes, the above condition is modified to [77]:

$$h^a_c(\sigma^2_{,a} - \omega^2_{,a}) = 0 \tag{3.50}$$

This must be satisfied if there are to be no density waves. i.e. the spatial gradients of the scalar shear and vorticity must coincide *everywhere*. From eq. (3.50) the irrotational silent universes, with $\omega^2 \equiv 0$, *always* have density waves unless they are homogeneous and isotropic, i.e. exactly FLRW. Since shear has scalar, vector and tensor contributions, while the vorticity has only vector contributions, this condition is never expected to be satisfied in a vortical dust universe, and hence we expect density waves to be generic in this case, as well as in the case involving pressure, relevant to cluster dynamics and structure formation. Vorticity and pressure fluctuations will further lead to changes in the propagation speed of these density waves, which are determined by the time evolution of the density gradient and the Hessian matrix of spatial derivatives of the density (see [77]).

The comoving density peak scenario of structure formation, with no density waves, is thus seen to be rather unlikely in realistic one fluid models. The effects of several fluids on the nature of density waves is unknown as yet, but may be expected to have bearing on the spatial variation of bias and the segregation of peaks in the baryonic and dark matter components [68], [83]. Finally we note that vorticity in the nonlinear multi-fluid case has not been studied. Given the importance of shear in the dynamics of segregation of baryonic and dark matter peaks [83], one may expect vorticity, if it is at all dynamically important at proto-galactic scales, to be similarly important for determining bias factors, through dynamical friction especially in bottom-up theories of structure formation where small nonlinear structures form first and merge to form galaxies. As mentioned after eq. (3.21), multi- fluid models involve non-zero fluxes, which will act as sources for vorticity [76].

3.12 Conclusions

The main results, ideas and conjectures of this chapter are summarised as follows:

1. The irrotational silent universes are unstable. Their lack of vorticity is nongeneric due to a coupling of the rotational mode to the density contrast in the quasi-linear regime. Since vorticity is known to exist even on galaxy cluster scales, where dissipative gas dynamics

are important, an irrotational model appears unrealistic for studies of non-linear collapse. The fact that the vorticity diverges with the density in strong collapse suggests that it is important dynamically. It may affect the natural configuration of gravitational collapse - spindle versus pancake.

2. The silent nature of the field equations is retained if one includes vorticity. Thus the fully nonlinear relativistic evolution is reduced to a coupled set of ordinary differential equations. However, it is likely, that the shear, vorticity and electric part of the Weyl tensor cannot be simultaneously diagonalised. The full integrability and consistency conditions for the VSU's have not been checked, and given the work of [80], it is possible that no fully consistent VSU's actually exist.

3. The Cauchy problem is, significantly more complicated due to the nonexistence of u^a-orthogonal, smooth cosmic time surfaces when the *vorticity* is in the quasi-linear or non-linear regime. This is intimately tied up with the averaging problem. By averaging over large enough scales, the vorticity is expected to vanish. However, the operation of averaging does not commute with the field equations.

4. Although the vorticity couples to the density and hence vortical knots (caustics) form in the same places as density knots, the spatial (Fourier space) variation of the vorticity is unknown, and is expected to backreact on the nonlinearity scale of the density. This should be studied to either decide whether it is important or not on galactic cluster scales today.

5. Fluid Vorticity may have an observable effect on the propagation of light bundles in astrophysical situations. This is a qualitatively new effect, and may be important in altering strong lensing statistics which are being used to constrain cosmogonic models [78].

6. Density waves, generic in the irrotational silent universes, are shown to be generic, with a more intricate existence structure, in both vortical silent universes, and the general nonlinear scenario with sound waves contributing as a source of density waves. Density waves are known to affect the nature of bias and segregation of peaks in the baryonic and dark matter components of the cosmic fluid [68].

Type		Conditions
I	–	*none*
II	–	$\eta^{(1)} = \eta^{(2)}$
III	–	$\eta^{(1)} = \eta^{(2)} = \eta^{(3)}$
D	–	$\eta^{(1)} = \eta^{(2)}, \eta^{(3)} = \eta^{(4)}$
N	–	$\eta^{(1)} = \eta^{(2)} = \eta^{(3)} = \eta^{(4)}$
O	–	$\Psi = 0$

Table 3.1: The different Petrov special types of the Weyl tensor in terms of the principal spinors of the Weyl spinor. If two or more spinors are proportional to each other they can always be made to be equal, as is the case quoted above.

Part II

The Cosmic Microwave Background and Gravitational Lensing

Chapter 4

Overview of The Cosmic Microwave Background

"To see a world in a grain of sand
And a heaven in a wild flower,
Hold infinity in the palm of your hand,
And eternity in an hour."

William Blake

4.1 Introduction

It is not exaggeration to say that the detection of temperature anisotropies in the Cosmic Microwave Background (CMB) radiation is to date probably the single largest step towards making cosmology a true science, in the sense that the predict-experiment-verify test cycle occurs with a short enough period to discourage theories from gaining religious-like weight.

This chapter aims to provide a self-contained introduction to both the current literature on the CMB and to the rest of this thesis. The former is a huge task as there have been literally hundreds of papers published on the subject during the 1990's. It is tempting therefore to try to tackle too much in such an overview. At the same time, it would be a gross mistake to concentrate on one's favourite model exclusively. Getting the best balance is neither time- independent nor an objectifiable 'fixed point". In fact the nature of the field will probably render it out of date within a couple of years. As a result, I have chosen to highlight many of the ideas which in my view will become important in the next decade as a way of obtaining detailed cosmological information. In essence this means moving to smaller angular scales and including CMB polarisation and spectral information.

As an outline of the chapter, we start with a discussion of CMB statistics, followed by sources of large angle temperature anisotropies: the Sachs Wolfe (SW) and Integrated Sachs Wolfe (ISW) effects. Small angle anisotropies are next, including the Doppler peaks, decoupling, contamination and the status of present and future experiments. We then argue the case for distinguishing between archetypical inflationary and topological defect models using the CMB. Moving away from temperature anisotropies to spectral distortions and then CMB polarisation, we explore information that will hopefully be extracted from the CMB in the next couple of decades. Finally we investigate the relationship between ergodicity, Gaussianity and the Copernican principle and the testing of the Copernican principle using the CMB.

This chapter predominantly chooses one formalism: the currently favoured one which dates back to Sachs and Wolfe 1966 [95]. The formalism has subsequently been cast in the form of Bardeen's gauge-invariant variables.

Because it is such a big field, there are a number of books which deal at least partially with the CMB, such as those by Partridge, Peebles, Padmanabhan etc... Further, it is almost impossible to divorce a study of the CMB from studies of structure-seeding and formation, topics which are barely touched on in this thesis. It will be assumed that the reader is familiar with at least the fundaments of modern structure formation theory, which can be found, e.g. in [148].

4.2 CMB Statistics

It is an anomaly, that the most widely recognisable feature of the CMB, the beautiful all-sky temperature maps, (see figure 4.1) are rarely used directly in actual analysis. Instead, dry, statistical, techniques are used, together with a wide variety of notations, all of which makes it a difficult field to break into. The initiation in my experience often unfortunately removes the initial joy of studying the subject, but is essential for a proper treatment of the subject.

In particular, the use of spherical harmonics is fairly universal, together with the assumptions of Gaussianity and ergodicity.

4.2.1 Ergodicity

A dynamical system is defined to be ergodic if its flow comes arbitrarily close to every point in the phase space in the large time limit $t \to \infty$.

A crucial result of ergodicity is that for an ergodic random field δ, spatial and ensemble averaging are equivalent. The ensemble average of δ at a point x^i is denoted $\langle\delta(x^i)\rangle$, and is simply the expectation value of the random variable $\delta(x^i)$. For an ergodic field in 3-d *flat space* this implies that for all points $x^{i'}$:

$$\langle\delta(x^{i'})\rangle = lim_{R\to\infty}(\frac{3}{4\pi R^3})\int_{x_i x^i < R^2} \delta_*(x^i)d^3x^i \tag{4.1}$$

where $\delta_*(x^i)$ denotes a specific realisation of δ. This equation simply tells us that we can substitute ensemble averages by averages over large volumes. Obviously this is a vital component to actually calculating statistical quantities in cosmology as we have only one universe to look at and are effectively limited to one view of the universe. As such we see a strong link with the cosmological principle which we will discuss later.

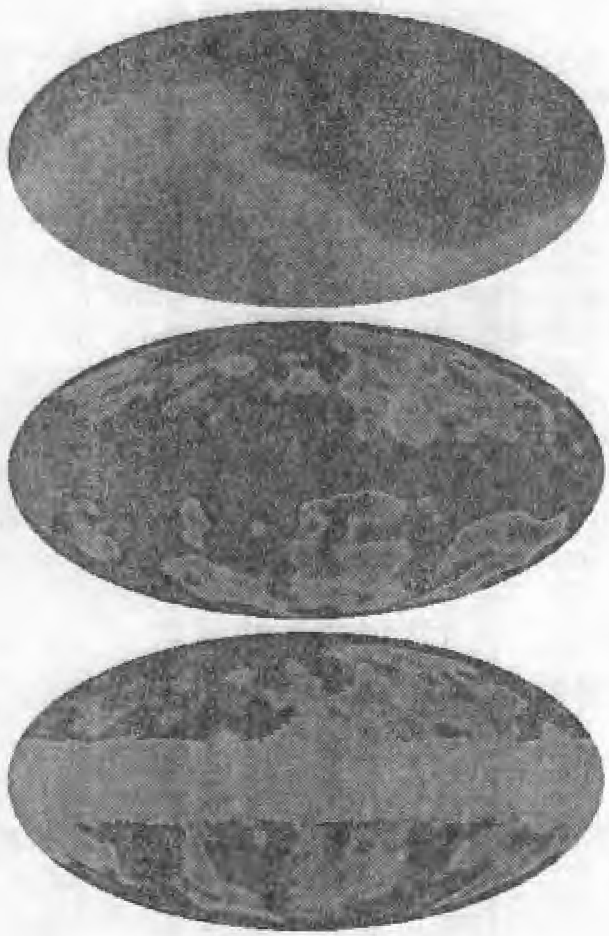

Figure 4.1: The all-sky maps from the 4-year data of the COBE satellite. Top: the raw data with dipole and galaxy contributions remaining. Middle: the dipole moments subtracted out. Bottom: the believed cosmological signal: the galaxy cut (galactic latitudes $b < 20^\circ$) has been employed. COBE averages over horns of 7° which can be seen in the lack of any small scale temperature variations. For comparison, the horizon size at decoupling subtends an angle of about 1° in the CMB today (if $\Omega = 1$). Picture courtesy of NASA.

4.2.2 Gaussianity

A random field (distribution) is Gaussian if it can be completely characterised by its two-point correlation function, with all higher moments of the distribution (correlation functions) either being zero (odd moments) or expressible in terms of the two-point correlation function (even moments). If further, the random field is statistically isotropic and homogeneous (the correlation function is invariant under rotations and translations) on a non-compact space, then we are sure that the resulting field is ergodic too [99].

In the case of the celestial sphere, S^2, which is compact, we cannot take the $R \to \infty$ limit required by eq. (4.1). Thus, even if our field is an isotropic, homogeneous Gaussian random field, our volume averages cannot tend to the required infinity. Therefore, when we use volume (or essentially solid angle) averages on the sphere, they will not correspond exactly to ensemble averages. This becomes severe for the low multipoles - particularly the quadrupole. This uncertainty is known as "cosmic variance", and because of it, one cannot "know" the low order harmonics of the CMB to better than about 20%, even with perfect instrumentation, since we have only one view of the CMB and only one universe. Here "know" refers to our inability to assign much statistical weight to the observed multipoles. Indeed, it is closely related to the Copernican principle, as discussed further at the end of this chapter.

4.2.3 The Two Point Correlation Function

In general, a basis for fields on the sphere is given by spherical harmonics $Y_{\ell m}(\theta, \phi)$, with coefficients $a_{\ell,m}$. However, as stated above, Gaussian fields are completely characterised by their two-point angular correlation function. The angular two-point correlation function on the sky is defined via:

$$C(\alpha) = \langle \frac{\Delta T}{T}(\theta) \frac{\Delta T}{T}(\theta + \alpha) \rangle \tag{4.2}$$

where the $\langle ... \rangle$ denote an ensemble average over all CMB skies. Obviously this is not possible for us to do (see later discussions of the almost-Copernican principle). Instead one makes the assumption of ergodicity. This allows one to approximately replace the ensemble average by single-sky volume averages, which one can perform as discussed in the section on ergodicity. One can then perform a standard decomposition of the 2-pt correlation function over Legendre polynomials:

$$C(\alpha) = \frac{1}{4\pi} \sum_{\ell=1}^{\infty} (2\ell + 1) W_\ell C_\ell P_\ell(cos(\alpha)) \tag{4.3}$$

where W_ℓ denotes the window function of the instrument which often has a Gaussian functional form itself. It is standard in the literature to present results for the angular spectrum in terms of $\ell(\ell + 1)C_\ell$ since for $\ell < 20$, flat $n = 1$ power spectra models predict this to be independent of ℓ, see figure (4.2).

Note that the spherical harmonics are simply a basis for fields on S^2 and hence require *no* assumptions for their implications. The two-point correlation function requires a yet untested assumption if it is to completely characterise the temperature fluctuations - that of Gaussianity.

We now leave the complicated subject of CMB statistics, which is a very active field at present, especially regarding optimal extraction of the true cosmic signal out of data which has both foreground and instrumental contamination [108].

4.3 Anisotropies at Large angular scales

When looking at a real temperature map of the CMB, one must always consider that it was produced from an instrument with finite resolution. The DMR instrument on COBE (Cosmic Background Explorer) for example, had a resolution of (beam width of) about 7°. At first glance the size of the averaging might not seem overly important. However, it encodes an important conceptual point. As the resolution of the instrument improves there occurs a critical transition when one suddenly starts observing anisotropies due to causal processes.

In the standard $\Omega = 1$ model the horizon size at decoupling occupies an angular scale of about 1° on the sky today. Therefore if one studies correlations on angular scales larger than this, one is not looking at anisotropies created by causal processes. They are either due to acausal effects or to initial conditions. By contrast, on smaller angular scales, the anisotropies may be due to driven acoustic oscillations, reprocessing of photons in clusters of galaxies, or to the evolution of the gravitational potential in the non-linear regime.

Photons traveling through spacetime will be redshifted such that their received frequency, ν_R, is:

$$\frac{\nu_R}{\nu_E} = \frac{1}{1+z} \tag{4.4}$$

where ν_E is the emitted frequency. If we treat the CMB spectrum as Planckian (see section 4.7 for a discussion of spectral distortions), the photon number is preserved and the occupation number of photons is:

$$n(\nu) = \frac{1}{(e^{h\nu/kT} - 1)} \tag{4.5}$$

By putting $n(\nu_E) = n(\nu_R)$ and substituting into the above equation from (4.4) it follows immediately that

$$\frac{T_R}{T_E} = \frac{1}{1+z} \tag{4.6}$$

Now in studies of the CMB we are interested in angular anisotropies. In particular those of the photon temperature field. We therefore are interested in variations of the photon temperature in the surface of last scattering (SLS) and we have used ν_E and T_E implicitly assuming that there is a well-defined, smooth surface from which photons were last scattered. However there are several subtleties with this. Firstly, how do we define the SLS ? Secondly, how do we take into account the effects of caustics from gravitational lensing which introduce non-differentiable curves in the SLS ? Finally, the SLS is not a surface but has a significant depth in redshift space, as will be discussed in the later subsection on recombination.

Now if we define an operator δ which operates in the space of our past null cone and which for different spacelike surfaces intersecting our past null cone, measures the difference between quantities (e.g. temperature) on the same two geodesics, we get from eq. 4.6:

$$\frac{\delta T_R}{T_R} = \frac{\delta T_E}{T_E} - \frac{\delta z}{1+z} \tag{4.7}$$

This tells us that the temperature measured across the sky at reception is due to intrinsic temperature variations in the SLS and to variations in the redshift-depth of the SLS. Both of these depend critically on the definition of the SLS.

Finally, although we made the assumption that the spectrum was Planckian (i.e. exactly black-body) to derive eq. 4.7, this is overly restrictive. We will derive the same equation if we allow chemical potential distortions of the spectrum which are independent of the photon frequency so that $\mu_R = \mu_E$ (see section 4.7).

4.3.1 The Sachs-Wolfe effect

One of the predictions of GR is that the frequency of light moving away from a massive body will be gravitationally redshifted, appearing cooler than it would had there been no background object. Now if we put a patchwork of linear overdensities and voids across the sky and propagate photons through it, we will get a combination of red- and blue-shifting, depending on whether the photons went through over- or under-densities. This is the physical origin of the Sachs- Wolfe (SW) effect, since the CMB photons had to move through the perturbations that seeded the galaxies and clusters we observe now at low redshifts.

As such we can see that as long as the gravitational potential (or more correctly the gauge-invariant potential Φ) does not evolve with time, then we should have a change in the energy (and hence frequency and hence temperature) of photons *on a single geodesic* which is just proportional to the difference in Φ between emission and reception.

$$\frac{\delta T}{T} = \frac{1}{3}\delta\Phi \tag{4.8}$$

where $\delta\Phi$ is the fluctuation in potential due to perturbations at the event where the geodesic intersects the surface of last scattering (treated as a $3-d$ hypersurface). [1] Note that if T is defined to be a temperature in a background FLRW model, then this is *not gauge-invariant* because by changing the FLRW background one will change the magnitude of δT. However, if one defines it to be the all-sky averaged temperature - i.e. $T = 2.73K$, then this is gauge-invariant.

In a flat universe with adiabatic perturbations and $\Lambda = 0$, this is the dominant source of anisotropy for both scalar and tensor modes on large angular scales, $\ell < 40$, if one neglects the decaying mode.

4.3.2 The Integrated Sachs-Wolfe effect

Since the evolution equation for the density contrast is second order there are two linearly independent solutions in general (see e.g. [16], referred to as the growing and decaying

[1] We can see already that one of the effects of gravitational lensing will be to change this intersection point with the surface of last scattering, thus causing a change to the overall temperature pattern.

modes. It is easy to show that the growing mode is actually constant and doesn't grow in a single-fluid dust $\Omega = 1$ universe. This is the reason that most of the anisotropy in a flat universe comes from the simple SW effect. However, in a open or Λ-dominated universe, this is not true. Intuitively, because of the added kinetic energy of the universe (we use here Newtonian cosmology for descriptive purposes) when $\Omega < 1$, *linear* perturbations stop growing when $z \simeq 1/\Omega$. Because of this the gravitational potential decays in time, and hence there is a $\dot{\Phi}$ term to be included as a source of the temperature anisotropy: the so-called Rees-Sciama effect, or since it was actually included in the original Sachs-Wolfe paper [95], the integrated-Sachs-Wolfe (ISW) effect. In this thesis we will refer to such anisotropy from global cosmological effects ($\Omega < 1, \Lambda \neq 0$) as the ISW effect, while the anisotropy due to local nonlinear matter evolution as the Rees-Sciama effect.

But why is there anisotropy associated with evolution of the gravitational potential ? It is simply because while crossing a perturbation, the potential evolves and hence the difference in potential at two points A,B on the photon path is different when the photon is at A and when it is at B. An everyday analogy of this is a roller coaster: usually the tracks do not move, so that ignoring friction, the end velocity is just dependent on the difference in potential at the start and end. However, if the height of the tracks changes with time (and hence also the potential), this will obviously not be true - one must integrate the rate of change of the height of the tracks over the whole path to account for the additional energy input or extraction. This integrated effect is just the ISW effect.

The total anisotropy along one geodesic, whose direction is given by the unit vector $\mathbf{n}$, is given by: [103, 115]

$$\left(\frac{\delta T}{T}\right)_R = \left(\frac{\delta T}{T}\right)_E + \delta\Phi + 2\int_E^R \frac{\partial\Phi(\tau,\mathbf{x})}{\partial\tau} + \mathbf{n}\cdot(\mathbf{v}_R - \mathbf{v}_E) \tag{4.9}$$

where $\mathbf{v}$ is the peculiar 3-velocity of the matter either at the intersection point of the geodesic with the surface of last scattering, or at reception. The difference is performed by parallel transporting the velocity vectors to the same spacetime point. The last term is the origin of the name "Doppler peak", but is in fact not the dominant cause of the small scale anisotropy. To calculate the integral we need to know the time evolution of Φ in the cases of interest. For dust in an $\Omega < 1$ universe, the longest lived mode (the "growing" mode) has a time evolution given by [16]:

$$F(\tau) = 5\frac{\sinh^2\tau - 3\tau\sinh\tau + 4\cosh\tau - 4}{(\cosh\tau - 1)^3} \tag{4.10}$$

where $\tau = 2(1-\Omega)^{1/2}$ showing that in this case we can use Ω as a time coordinate if we so wish.

From this equation one can calculate the anisotropy due to the ISW effect. It has been found [103, 104], that for low-Ω (i.e. $\Omega < 0.4$) models the ISW effect dominates the standard SW effect for adiabatic perturbations on large angular scales. Since $\dot{\Phi} < 1$, we see that the contribution is of the opposite sign to the usual SW effect, so that roughly a hot spot due to the SW effect corresponds to a cold spot of the ISW effect. Further, we see from figure (4.3) that $\dot{\Phi}$ will be most negative as $\tau \to 2$ (i.e. small redshift). Therefore the majority of ISW anisotropy, which is the area above the $\dot{\Phi}$ curve, is due to the freezing out of linear

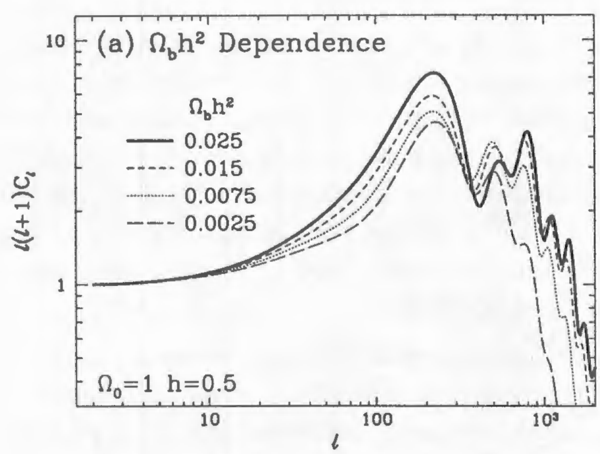

Figure 4.2: The primary angular spectrum for a standard inflationary, CDM model. Note the flat large-angle plateau (LAP) for $\ell < 20$ and the Doppler peaks for $\ell > 100$. Silk damping due to the finite thickness of the last-scattering surface damps out anisotropy for $\ell > 1500$. Secondary anisotropies will mainly contribute in this region. Shown is the effect of changing the baryon content of the universe. Decreasing Ω_b reduces the height of the Doppler peaks and increases the effectiveness of Silk damping due to the increase in the thickness of the surface of last scattering. From [100].

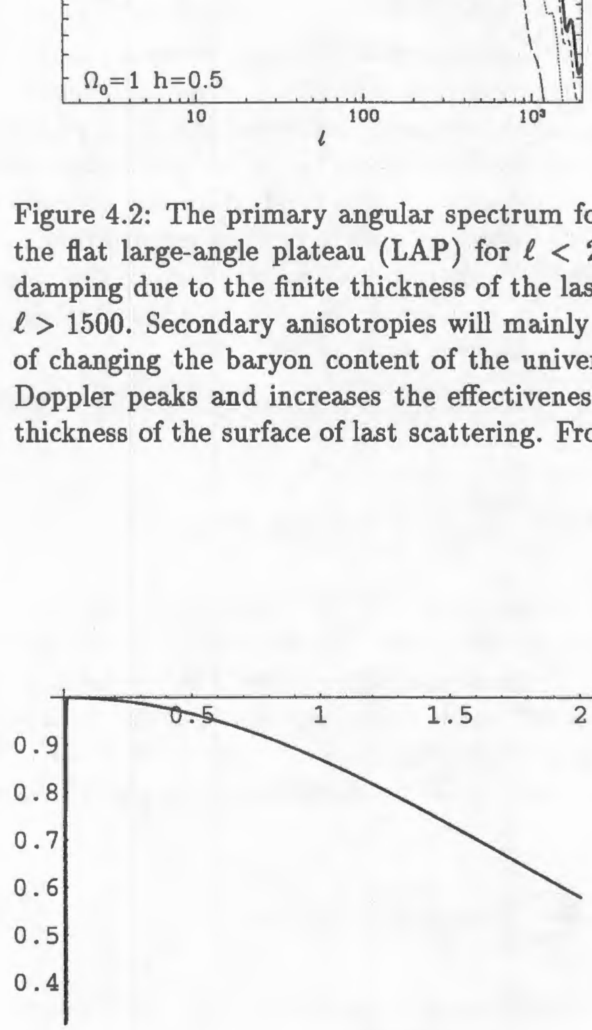

Figure 4.3: Φ vs τ. The decay of the "growing mode" of the gauge-invariant gravitational potential, Φ, in an open dust universe. The time coordinate is $\tau = 2(1-\Omega)^{1/2}$. The big bang corresponds to $\tau = 0$ and future time-like infinity to $\tau = 2$.

perturbations at low redshifts, $z \leq 1/\Omega$, and not due to anything happening on the surface of last scattering.

In the case of adiabatic perturbations, there is no perturbation in the total equation of state. Therefore in regions where there are matter overdensities, the photon number density must also increase. Thus we expect to be able to write the variation in the photon energy density on the surface of last scattering in terms of the fluctuations in the potential Φ. This is indeed true and using the fact that

$$\rho_\gamma = aT^4 \longrightarrow \frac{\delta T}{T}|_E = \frac{\delta \rho_\gamma}{4\rho_\gamma}|_E$$

one finds that $\delta\rho_\gamma/4\rho_\gamma|_E = -2/3\delta\Phi$, so that matter contributions to the SW effect are partially cancelled by the intrinsic fluctuations in the photon temperature. In this case the total anisotropy is given by:

$$\left(\frac{\delta T}{T}\right)_R = +\frac{\delta\Phi}{3} + 2\int_E^R \frac{\partial\Phi(\tau,\mathbf{x})}{\partial\tau} + \mathbf{n}\cdot(\mathbf{v}_R - \mathbf{v}_E) \tag{4.11}$$

Note an important point however. When we wrote down that $\frac{\delta T}{T}|_E = \frac{\delta\rho_\gamma}{4\rho_\gamma}|_E$, we did not specify where or what the variation δ was. In particular we assumed that in the hypersurface of last scattering, $\delta\rho_\gamma \neq 0$. However one is not sure that this is true. Consider the following argument.

The surface of last scattering is usually defined as the hypersurface of unit optical depth, which to first order is due to Thompson scattering, since Rayleigh scattering etc.. are much less important at these epochs. The optical depth for Thompson scattering is defined as:

$$\tau = \int \sigma_T n_e x_e S' d\tau \tag{4.12}$$

where S is the scale factor, τ the conformal time, $' = d/d\tau$, and n_e, x_e the electron density and ionisation fraction respectively. σ_T is the cross section for Thompson scattering. Now if we neglect the backreaction of density perturbations on the ionisation fraction, which will presumably cause second order effects, the optical depth is dependent on the free electron density,[2] which in the tight coupling limit will have the same spatial distribution as the photons, which in turn for adiabatic perturbations trace the overall matter potential fluctuations. Add to this the fact that the isodensity surfaces for the free electrons and photons coincide, so that the photon energy density is constant on the surface of last scattering, and since $T_\gamma \propto (\rho_\gamma^{1/4})$, there are no photon temperature fluctuations, $\delta T|_E = 0$. But since the adiabatic assumption is that the photon and matter perturbations are proportional, $\delta T = 0 \Rightarrow \delta\phi = 0$ [107], and there is only a SW contribution from the velocity term. We will discuss the more realistic situation of non-instantaneous recombination in the later section on decoupling.

Finally we note that there is another case in which the ISW effect contributes to the CMB anisotropy: the so-called early-ISW effect (in contrast with the late, low-redshift ISW effect discussed so far). This occurs in realistic 2-fluid models because, near the change from radiation to matter domination, the gravitational potential evolves in time, even if the background geometry is flat [102]. This typically contributes at smaller angles [104].

[2] We have ignored here the nonlinear agreggations of free electrons at low redshifts in cluster cores. However, they occupy a relatively small fraction of the sky.

4.3.3 Heresy - decaying modes

It is completely standard practice to drop the decaying mode from any calculations involving the CMB or structure formation. This is based mainly on inflation but also on the belief that present large - scale structure evolved from linear perturbations, seeded at $z \gg 1000$. If this is true then neglecting the decaying mode is fully justified. However, consider the following "nightmare" scenario:

Some mechanism, say topological defects, creates a spectrum of perturbations around decoupling with a great deal more power in the decaying mode than in the growing mode. The growing mode evolves to form the present large-scale structure, while the decaying mode decays quickly around the epoch of decoupling and plays no real role in structure formation. However, because of the extremely rapid death of the decaying mode [16], $\dot{\Phi} \neq 0$, there is a significant early ISW contribution which should dominate the CMB anisotropy, depending on the details of the decaying mode spectrum. In fact by tuning the decaying mode spectrum we can have it dominate any part of the angular spectrum we want ! In particular, we can tune the spectrum such that a CDM model could be normalised to large-scale structure observations (e.g. σ_8, the mass variance within spheres of 8 Mpc), and simultaneously match it to the COBE normalisation, thus solving the problem that CDM produces too much power on small scales when normalised to COBE. Of course this is nothing overly special since we have essentially just introduced another free parameter - but it is a very natural one, that cannot at present be excluded except by theoretical prejudice.

4.3.4 Null cone calculations

The considerations above regarding the dropping of the decaying mode and the vorticity in CMB calculations, may prompt one to look for calculations which do not make a splitting into growing and decaying and scalar, vector & tensor modes. Such a formalism exists, and is rather beautiful, but has been relatively little explored [207]. It relies on observational coordinates: $\{w, y, \theta, \phi\}$. Here θ and ϕ are angular coordinates on the sky and w, the time-coordinate, specifies a past null cone completely. y is the radial coordinate down each null cone and can be chosen to be area distance, redshift, or similar quantity.

We will not however, examine the formalism in any depth here, since it uses the rather technical fluid null tetrad formalism. Indeed the remaining outstanding feature of this formalism is the relation of the anisotropy calculated in these coordinates to actual matter properties: i.e. the power spectrum, two-point correlation function etc...

We will however discuss another recent innovation which has also been labeled as being of "null-cone" - type. However, it does not use the observational coordinates which are adapted to the actual null cone. Rather the term comes from a special decoupling that occurs which allows a significant decrease in computation time when implemented numerically.

In this approach [112], the coupled Boltzmann hierarchy of equations (after angular and momentum integration) is solved formally as a pair of integral equations (for each wave number k of purely scalar perturbations):

$$\Delta_T = \int_0^{\tau_0} d\tau e^{ik\mu(\tau-\tau_0)}\, e^{-\kappa}\left[\dot{\kappa}e^{-\kappa}(\Delta T_0 + i\mu v_b + \frac{1}{2}P_2(\mu)\Pi) + \dot{\phi} - ik\mu\dot{\psi}\right] \tag{4.13}$$

$$\Delta_P = -\frac{1}{2}\int_0^{\tau_0} e^{ik\mu(\tau-\tau_0)}\; e^{-\kappa}\dot{\kappa}\,[1 - P_2(\mu)]\,\Pi\; d\tau \tag{4.14}$$

where ϕ, ψ are the Bardeen potentials, v_b is the velocity of the baryons, Δ_T is the anisotropy due to the single mode k in a given direction. By defining it to be the variation in temperature from the all-sky mean temperature it is made gauge-invariant. Δ_P is the polarisation (see section 4.8) of the photons gained from the mode k. $\mu = \cos\theta$ is the cosine of the angle between the wave vector, $\mathbf{k}$, and line of sight and is the only angular dependence after integrating axially about the wave vector $\mathbf{k}$, describing the plane wave perturbation. Π is the scalar anisotropic stress [104] and κ is the optical depth from Thompson scattering. $P_2(\mu)$ is the second Legendre polynomial. These integrals can be further simplified by integrating by parts and dropping the boundary terms because they only affect the monopole. This allows Δ_T, Δ_P to be written as:

$$\begin{aligned}
\Delta_{T,P} &= \int_0^{\tau_0} e^{ik\mu(\tau-\tau_0)}\; S_{T,P}(k,\tau)\; d\tau && (4.15)\\
S_T(k,\tau) &= g\left[\Delta_{T_0} + \psi - \frac{\dot{v}_b}{k} - \frac{\Pi}{4} - \frac{3\ddot{\Pi}}{4k^2}\right] && (4.16)\\
&+ e^{-\kappa}(\dot{\phi} + \dot{\psi}) - \dot{g}\left[\frac{v_b}{k} + \frac{\dot{\Pi}}{4k^2}\right] - \frac{3\ddot{g}\Pi}{4k^2} &&\\
S_P(k,\tau) &= -\frac{3}{4k^2}\left[g(k^2\Pi + \ddot{\Pi}) + 2\dot{g}\;\dot{\Pi} + \ddot{g}\Pi\right] && (4.17)\\
& && (4.18)
\end{aligned}$$

where $g = \dot{\kappa}e^{-\kappa}$ is the visibility function. Compared with the observational coordinates null cone calculation referred to earlier in this section [207] this has both advantages and disadvantages. Firstly, it is rather complex, it makes a splitting into scalar, vector and tensor modes and it needs to be integrated over k to give the final anisotropy and polarisations. It's advantage is that it allows one to identify the origin of the anisotropies: the first term of S_T is the intrinsic anisotropy from non-adiabatic perturbations, the second term is the SW contribution while the third and eighth terms are velocity contributions. The term $e^{-\kappa}(\dot{\phi}+\dot{\psi})$ is the (early + late) ISW effect while the anisotropic stresses Π which, together with the visibility function, is the way polarisation couples to the temperature anisotropy. In the approximation of tight-coupling (i.e. the mean-free photon path is zero), and instantaneous recombination, $g = \delta(\tau - \tau_{recomb})$, Π can be neglected and there is no polarisation induced. Conversely, we can already see that if there is reionisation, there will be extra contributions to the polarisation due to the extra width of g in time.

To obtain the angular power spectrum one must expand in eigenfunctions (plane waves in the flat geometry case) and perform ensemble averages with integration over μ. This requires $\Delta_{T,P\;\ell}$, which are given in the flat geometry case by:

$$\Delta_{T,P\;\ell}(k,\tau = \tau_0) = \int_0^{\tau_0} S_{T,P}(k,\tau)\; j_\ell[k(\tau - \tau_0)]d\tau \tag{4.19}$$

where $j_\ell(x)$ is the spherical Bessel function. This is where the main advantage of this formulation lies: this integral is a function of ℓ and of specific cosmological model. But these dependencies are separated: the spherical Bessel function carries all of the ℓ-dependence, and is independent of specific cosmological model, while $S_{T,P}$ carries the information about the temperature or polarisation sources in terms of the specific model characteristics. This means that one can calculate the $j_\ell(x)$ once and then use it in the integral. Standard formulations which solve the coupled hierarchy of Boltzmann equations do not have this neat splitting which follows from the use of a null-cone formalism. This leads to a reduction in calculation time that can be as large as two orders of magnitude [112].

4.4 Small angular scales

4.4.1 Doppler peaks

The angular power spectrum for standard cosmogonies involve peaks and troughs known as the "Doppler peaks" which are found in the C_ℓ vs ℓ plots at $\ell \geq 100$ (see figure 4.2). More correctly they might perhaps be known as Sakharov peaks or acoustic oscillations, partly because the peaks have little to do with the Doppler effect associated with the motion of the surface of last scattering. However, the name is popular and entrenched, so we will also use it.

4.4.2 An intuitive insight into the Doppler peaks

A full calculation of the small scale anisotropy requires that we drop all of the simplifying assumptions that are reasonable in other circumstances. These involve:

(1) Non-equilibrium thermodynamics,

In the region of decoupling the strong coupling applicable during the early universe between the photons and electrons begins to break down and one cannot strictly speak of thermodynamic equilibrium.

(2) multiple, coupled fluids and the distinctions between adiabatic and isocurvature perturbations,

(3) The finite width of the surface of last scattering and the atomic physics of recombination,

(4) Foreground contamination effects from the galaxy and beyond.

Here we will only discuss the last two points, and even then only schematically.

4.4.3 Recombination and decoupling

As mentioned in the last section, a detailed study of the physics (as opposed to relativistic dynamics) involved during decoupling and recombination is required for a complete understanding of the small scale CMB spectrum. Now the terms "decoupling" and "recombination" describe very different physical processes but are used almost synonymously in all but the most technical discussions of the CMB. This is essentially due to the dominance of Thompson scattering in photon-electron interactions at temperatures of the order of $3000K$ over other processes such as Rayleigh scattering [96]. Hence when the recombination occurs, Thompson scattering becomes negligible and there is very little coupling between the photons and matter. However, we will note several points where there are subtle differences between the two.

The discussion that follows is essentially that of Peebles (1968) [97] and Ma and Bertschinger (1995) [98]. Now in a general, accurate calculation of anisotropy formation at small scales one needs to know the Thompson scattering cross section accurately. To achieve this in turn means that we need to know accurately the free electron density, $n_e(\tau)$, as a function of (conformal) time, τ. In this section we will derive the equation that gives it to us.

Since Helium has a much higher ionisation potential than hydrogen, the ambient photons fall below the temperature required to ionise helium long before hydrogen recombination and hence we treat the helium component as completely neutral.

When hydrogen recombines the previous ionisation equilibrium breaks down because of the lack of free electrons. One must numerically study the evolution of the ionisation fraction which then closes the set equations required for discussion of this epoch.

We define the ionisation fraction of hydrogen as $x_H \equiv n_e/n_H$ where n_H is the total number density of hydrogen nuclei. The rate of ionisation is governed by [97]:

$$\frac{dx_H}{d\tau} = SC_r\left[\beta(T_b)(1-x_H) - n_H\alpha^{(2)}(T_b)x_H^2\right] \tag{4.20}$$

where S is the scale factor, $\beta(T_b)$ is the collisional ionisation rate for hydrogen from the ground state:

$$\beta(T_b) = \left(\frac{m_e k_B T_b}{2\pi\hbar^2}\right)^{3/2} exp(-B_1/k_B T_b)\alpha^{(2)}(T_b) \tag{4.21}$$

where $B_1 = 13.6eV$ is the ground state binding energy and:

$$\alpha^{(2)}(T_b) = \frac{64\pi e^4}{(27\pi)^{1/2}m_e^2c^3}\left(\frac{k_B T_b}{B_1}\right)^{-1/2}\phi_2(T_b) \tag{4.22}$$

is the recombination rate to excited states, and

$$\phi_2(T_b) \simeq 0.448\ln\left(\frac{B_1}{k_B T_b}\right) \tag{4.23}$$

Now what is the quantity C_r ? This is related to the way in which recombination can actually occur. The simplest process would be:

$$e^- + p \rightarrow H + \gamma \tag{4.24}$$

with the hydrogen atom in the ground state. However, this is highly suppressed due to the large Lyman alpha and Lyman continuum opacities, i.e. there are lots of photons at just the right energies to reionise the newly formed atoms even though the temperature is falling rapidly below that needed for efficient ionisation in general. Therefore recombination must predominantly precede either via 2-photon decay from the 2s to the 1s level, or via redshifting of Lyman alpha photons away from the line centre due to the expansion of the universe. The first process occurs with a rate $\Lambda_{2s\rightarrow 1s} = 8.227s^{-1}$ while the second only becomes effective at the end of the recombination period. In fact, the ratio of the rates of these two processes is given by [97].

$$\frac{\text{Lyman}-\alpha\ \ \text{redshifting}}{2-\text{photon decay}} = 0.0022\frac{n_{1s}}{n}T_4^{-3/2} \tag{4.25}$$

where T_4 is the radiation temperature in units of $10^4 K$, n is the number of hydrogen atoms plus free electrons and n_{1s} is, as it suggests, the number of atoms in the $1s$ state - which for $T_b \ll 10^5 K$ is well approximated by $(1 - x_H)n_H$. Since this above ratio is very small when $T_4 \simeq 0.1 - 0.3$, the bulk of recombination must occur via the 2-photon decay ($2s \rightarrow 1s$) to the ground state.

Finally, the quantity C_r describes the reduction of the recombination rate due to reionisation of hydrogen atoms in which the electron is in the 2s state waiting to decay via the above 2-photon process. In fact, C_r is just the ratio of the net reionisation rate, dominated by 2-photon decay, to the sum of recombination and ionisation rates for the $n = 2$ level:

$$C_r = \frac{\Lambda_\alpha + \Lambda_{2s\rightarrow 1s}}{\Lambda_\alpha + \Lambda_{2s\rightarrow 1s} + \beta^{(2)}(T_b)} \tag{4.26}$$

Here $\Lambda_\alpha = \frac{8\pi \dot{S}}{S^2 \Lambda_\alpha^3 n_{1s}}$ is the rate of recombination due to redshifting of Lyman-alpha photons out of the line, $\beta^{(2)}(T_b) = \beta(T_b) e^{h\nu_\alpha / k_B T_b}$, and $\lambda_\alpha = \frac{8\pi\hbar c}{3B_1}$, with ν_α the corresponding frequency.

Thus we have "fully" described the evolution of the free electron density through recombination. Here we have ignored perturbations in the electron density, correlations in the plasma which will change the collision functional in the Boltzmann equation and resonance-line radiation. However, these are believed to be of lesser importance.

4.4.4 Foreground contamination

As one moves to smaller and smaller angular scales one is moving more and more into the regime of galactic physics and one therefore has to be concerned with contributions to the microwave background from non-cosmological sources such as dust, free-free emission and synchrotron radiation to mention just three. One is also faced with a more subtle point - the fact that there are true anisotropies due to gravitational effects from matter clustering but on non-cosmological scales and at low redshifts (e.g. the Rees-Sciama effect). These effects are generally known as secondary although in special cases they can be dominant.

Here we limit ourselves to a discussion of truly non-cosmological contributions that must be subtracted out to get the true signal. Our discussion will not be thorough but is instead rather qualitative aiming to give an overview of the experimental complexities involved.

4.4.5 Extragalactic foreground

We may split the known extragalactic contaminants into two main categories: radio point sources and infra-red point sources. Unfortunately we have no real understanding of the frequency dependence of the power from either of these sources although the angular power spectrum is known to be a flat white noise spectrum in both cases: $C_\ell = const.$ Since the power spectrum from cosmological origin is expected to behave like $C_\ell \propto \ell^{-\alpha}$ for some $\alpha > 0$ and large ℓ, one can see that the extragalactic contamination will dominate for small enough angular scales and cannot therefore be neglected and must be removed from the data by hand [113].

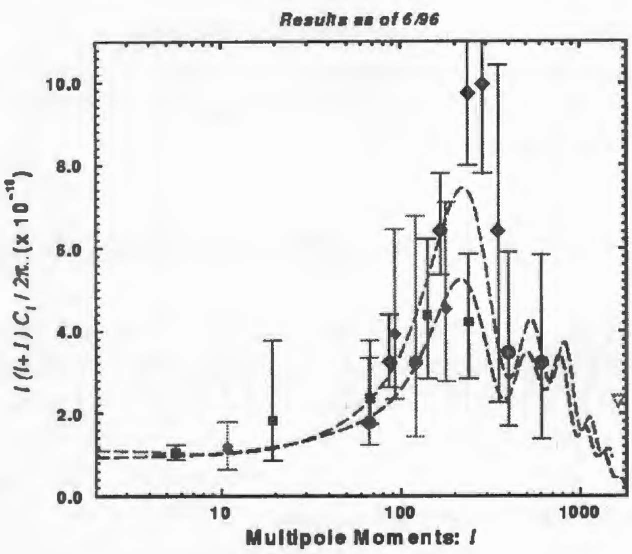

Figure 4.4: The experimental results as of mid-1996. The large errors bars especially at large-ℓ are the main hurdle to future progress. Superimposed are some typical inflationary spectra. From http://dept.physics.upenn.edu/ www/astro-cosmo/cmbr/pack.html.

4.4.6 Galactic foregrounds

Here we may divide the contamination into that due to the solar system and that due to extra-solar sources. The solar sources mainly impact on the design of the experiment, and where it will observe (or its spin axis and trajectory in the case of satellite experiments). One must avoid as much as possible the side-lobes due to the sun, moon and all other planets. In addition there is atmospheric emission, which is a great problem for balloon and ground-based instruments.

From galactic sources one must deal with dust, free-free and synchrotron emissions. The power spectrum of galactic dust has been calculated from the IRAS all-sky survey and from the COBE FIRAS & DIRBE data. The IRAS survey in particular provides excellent resolution data and shows that the overall shape of the spectrum is basically independent of galactic latitude with a typical dependence of $C_\ell \propto \ell^{-3}$ for $\ell > 100$.

4.4.7 Experimental results

At present there are several experiments that are probing the small scale CMB spectrum. However, as is evident from theoretical plots (e.g. fig 4.2), this is the region where the power drops significantly from cosmological sources and as discussed above, also the region in which contamination effects are most numerous, significant and difficult to subtract.

At present there are only upper limits on the anisotropy at really small scales, see figure (4.4). The WD, OVRO and ATCA experiments probe the spectrum between $\ell \simeq 700-5000$

and are not shown in the figure. However, this will be a region of great focus in the next two decades, simply because once the normalisation was fixed (modulo the approximately 20% statistical uncertainty due to cosmic variance) on the largest scales, the experiments with the most descriminative power are those at the smallest angular scales, due to a "lever-arm" effect when combined with the COBE data at $\ell \leq 20$.

4.5 Distinguishing between inflation and topological defects

The section on Doppler peaks assumed an adiabatic inflationary scenario which predicts a set of Doppler peaks whose heights and positions give us information about the parameters of the inflationary model and those of the FLRW background model. What if topological defects were responsible for seeding the primordial density fluctuations ? How would the CMB look in that case ? We have only to consider the cases of cosmic strings and global textures as possible scenarios. Monopoles and domain walls being excluded on the grounds that they would come to dominate the energy density of the universe at high redshift if they were responsible for seeding structure formation.

4.5.1 Non-gaussian features

The most important feature of topological defects is that they imply non-Gaussian features in the CMB at scales around the horizon size at decoupling (e.g. $1°$ if $\Omega = 1$). In the case of cosmic strings this is because of temperature discontinuities induced transverse to the length of the string while in the case of textures there are large radial gradients. Putting this statistically means that the phases of the temperature anisotropy spectrum are not uniformly distributed but are correlated. Any detection of a non-Gaussian signature, e.g. from a non-zero bispectrum or n - point correlation function ($n > 2$), would signal the existence of some seeding process beyond the standard model (inflation coupled to a CDM/MDM model).

Unfortunately, most of the statistics proposed so far do not yield a feasible test of Gaussianity because of cosmic variance. Physically this is because with only one realisation of the statistics, it is possible to have non-Gaussian signatures even if the parent distribution is truly Gaussian. [3] Allied to this, topological defect models do not predict non-Gaussian statistics that are significantly different from the cosmic variance limits. Fortunately the Doppler peaks offer at present, a fairly robust way of discriminating between the two fundamental theories.

4.5.2 The Doppler peaks

Both cosmic strings and textures are believed not to produce secondary Doppler peaks and that the first Doppler peak occurs at a much larger ℓ than in inflationary models [101], i.e. $\ell = 350 - 600$ compared with $\ell = 220$ if $\Omega \sim 1$. $\ell = 400$ is where one would expect the first Doppler peak in a very low density, $\Omega < 0.3$, open inflationary universe. But once Ω

[3] A simple example is provided by tossing coins. Toss a coin 3 times and it is not possible to distribute the outcomes equally between heads and tails. Toss 10^8 coins and the underlying parent distribution is clear and any deviation from a uniform distribution will be obvious.

is known, this, perhaps with the non-Gaussian features, seems to imply that topological defects should be falsifiable as a theory within the next decade or so [109], especially given the resolution of the future CMB satellites, MAP and COBRAS/SAMBA. However, as discussed in later chapters, the effect of caustics due to strong gravitational lensing may mimic some of the non-Gaussian features of topological defects and make detection of a defect signal more difficult (see chapter 7). This confusion in the source of the non-Gaussian signal is important since weak lensing may wipe out secondary Doppler peaks even if they existed at high z [126].

Note that all the treatments carried out so far have assumed that the vorticity was zero, and hence in the case of cosmic strings, have neglected the vorticity generated in bow-shock waves and wakes of the relativistic cosmic strings which would act as a vector source of CMB anisotropy, if non-negligible at decoupling.

4.6 Future, second-generation CMB anisotropy experiments

4.6.1 Space missions

The great success of COBE has emphasised the advantages that lie in going into space to do CMB experiments. At present there are two space missions that have been approved. These are COBRAS/SAMBA, and MAP. The main aim of these groups is to map the whole sky at resolutions much better than the 10° of COBE. A resolution of 10' is required, which when combined with the large scale 3-D galaxy surveys such as the Sloan Digital Sky Survey, and FIRAS-level measurements of the CMB spectral distortions at centimeter wavelengths, should answer many of our questions regarding the viability of theories such as inflation, topological defects, & models of dark matter, reionisation and the fundamental parameters of the universe: the baryon content, Ω, Λ and the Hubble constant.

COBRAS/SAMBA and MAP

We will consider these two satellite experiments in more detail. MAP was selected by NASA for its Medium Class Explorer program and although it has a lower angular resolution than COBRAS/SAMBA, will also provide polarisation information, a vital additional service.

The COBRAS/SAMBA experiment was selected by ESO and will measure the CMB in 8 frequency bands (as opposed to the 3 of COBE) in the range 30-800 GHz with peak sensitivity $\Delta T/T \simeq 10^{-6}$. It will map the angular spectrum from $\ell = 2 - 1500$ and will provide measurements of the Sunyaev-Zel'dovich effect for over 1000 rich clusters, which in conjunction with X-ray data, will provide an independent estimate of the Hubble constant and Ω.

With all experiments the increased resolution and required sensitivity will require an orbit far from the sun and earth. The COBRAS/SAMBA experiment will be located near L2, one of the earth-sun Lagrange points, with the spin axis pointed at the sun. It will be equipped with both bolometers and radiometers, enabling the large frequency coverage. Another very important difference between modern CMB experiments and COBE DMR is that they will only use a single telescope compared to the two horns used by DMR. Images of the whole sky will be built up over 6 months as it orbits the sun and because of the single beam nature of the experiment, data will be able to be analysed as it is downloaded from the satellite, again in contrast to the COBE data [105].

4.6.2 Ground-based and balloon experiments

Many of the current ground and balloon CMB experiments will carry on till the turn of the century. In addition there are several long-duration balloon flights planned for the near future such as BOOMERANG, TOPHAT and ACE.

TOPHAT is conceived of as a balloon with the detectors on top of the balloon, as opposed to the usual gondola configuration [115]. BOOMERANG is an extension of the current MAXIMA experiment in many ways, and was planned to make flights over Antarctica at the end of 1996. Finally ACE (Advanced Cosmic Explorer) is planned to be a much longer duration balloon experiment with flights lasting around three months instead of several days. With three flights it might map three quarters of the sky at a resolution of $10'$ and pending funding questions, could fly by the year 2000.

As far as ground-based experiments are concerned, only increases in detector technology are expected to give improvements, due to the fundamental limitations imposed by the atmosphere (which were minimised by going to the South Pole - SP89 - and high altitudes. - Tenerife - and using complex triple-beam chopping strategies). A caveat to this is the use of interferometers to map the anisotropy on small scales, such as the Cambridge CAT, with several more awaiting funding. They would be particularly useful when studying the Sunyaev-Zel'dovich effect in clusters, where in principle simultaneous measurements of the CMB polarisation could be made to get the peculiar velocities of the clusters (see sec. 4.8).

4.7 Spectral distortions

The publicity due to COBE's success shone very brightly on the Digital Microwave Radiometer (DMR) team lead by G.F. Smoot. The results warranted this. However the other two instruments on board - FIRAS and DIRBE - provided very high quality results which to a large extent have been overshadowed. We will focus here on the implications of the *Far InfraRed Absolute Spectrophotometer (FIRAS)* experiment on COBE. FIRAS was a polarising Michelson interferometer which has provided by far the most accurate constraints on the CMB spectrum to date.

From a cosmological point of view, knowledge about the spectral distortions of the CMB proved very effective at eliminating theories of structure formation, such as explosion models and many of the isocurvature (PBI) models. FIRAS further provided information on the processes occurring before decoupling, something DMR could not do, and information about the general Intergalactic medium. But how did it do this ?

4.7.1 The physics of spectral distortions

Before decoupling the photon gas that eventually formed the CMB was in local thermal equilibrium with the various species of particles which had not yet "frozen out". The photon distribution was well described by a Planck blackbody distribution because of this. Any phase transitions that released energy into the photon gas are not seen as distortions in the spectrum today, primarily due to photon producing processes such as radiative Compton scattering [114]

$$\gamma + e^- \to \gamma + \gamma + e^- \tag{4.27}$$

which take ultra-high energy photons and redistribute the energy allowing relaxation to a blackbody spectrum. Radiative Compton scattering can achieve this cleansing of the spectrum until a redshift $z \simeq 1.4 \times 10^6 (\Omega_b h^2)^{-0.4} \simeq 10^7$ for standard ranges of models. After this there is only non-radiative Compton scattering which is an adiabatic process in the sense that it preserves photon number. Further, one must use the correct quantum-mechanical treatment i.e. Bose-Einstein statistics, to describe the photon occupation number, which is given by:

$$\eta(x) = \frac{1}{e^{x+\mu(x)} - 1} \tag{4.28}$$

where

$$x \equiv \frac{h\nu}{kT_e} \tag{4.29}$$

is the dimensionless frequency. The distribution is now characterised by two numbers: the temperature and the chemical potential, μ. The chemical potential describes distortions from blackbody due to high-redshift interactions, namely non-radiative Compton scattering and Bremsstrahlung (free-free) creation of photons. These two processes compete and there exists a frequency, x_{CB}, at which the photons are Compton scattered (diffuse) to higher frequencies as fast as they are created (at lower frequencies) by the Bremsstrahlung process. However, in general we expect the spectral distortions to be positive at low frequencies (due to Bremsstrahlung) with the true spectrum lying above a blackbody of the same local temperature. As we cross x_{CB}, the two spectra cross and for higher frequencies the true spectrum lies below the blackbody one. For even higher frequencies the two spectra cross once again after which the true spectrum again lies above the blackbody because of the photons received from between the two crossing points due to (inverse) Compton scattering. Another way of stating this is that the spectrum in the Rayleigh-Jeans region is sparse of photons, with a deficit characterised by ΔT_{RG}:

$$\Delta T_{RG} \simeq -2yT_\gamma \tag{4.30}$$

while in the Wien (high frequency) region the spectrum is over populated with photons which diffused there from the Rayleigh-Jeans regime. The characterising parameter is the so-called Compton y-parameter:

$$y = \int_0^z \frac{k[T_e(z) - T_\gamma(z)]}{m_e c^2} \sigma_T n_e(z) c \frac{dt}{dz'} dz' \tag{4.31}$$

which for small y is directly related to the total energy transferred by Inverse Compton scattering:

$$\frac{\Delta E}{E} = e^{4y} - 1 \simeq 4y \tag{4.32}$$

The distortion of the spectrum today due to free-free (Bremsstrahlung) processes is given by:

$$\Delta T_{ff} = T_\gamma \frac{Y_{ff}}{x^2} \tag{4.33}$$

where T_γ is the undistorted photon temperature, x is again the dimensionless frequency and Y_{ff} is the optical depth for free-free emission [115]:

$$Y_{ff} = \int_0^z \frac{k[T_e(z) - T_\gamma(z)]}{T_e(z)} \frac{8\pi e^6 h^2 n_e^2 g}{3m_e(kT_\gamma)^3\sqrt{6\pi m_e k T_e}} \frac{dt}{dz'} dz' \tag{4.34}$$

where n_e is the electron density and g is the Gaunt factor [116]. Note that the electron and photon temperatures are not equal in general and note the similarities with the definition for y (eq. 4.31). What does this distortion look like ? It has a quadratic rise in temperature at low frequencies and is in fact the dominant signature for warm ($T_e \simeq 10^4$) plasmas after decoupling. Free-free emission attempts to thermalise the spectrum to the plasma temperature while Compton scattering thermalises to near-blackbody with a constant chemical potential μ_0:

$$\mu_0 \simeq 5.6y \tag{4.35}$$

To describe the two competing effects one must use a frequency-dependent chemical potential:

$$\mu(x) = \mu_0 \ \exp(-2x_{CB}/x) \tag{4.36}$$

This is a good phenomenological description, but where did this fundamental parameter y come from and why do we talk of Compton scattering as a diffusion process ? To answer these questions we must have an equation for the time and frequency variation of the spectrum under all the different processes. This partial differential equation is generally known as Kompaneets' equation [117] and requires a kinetic theory description for the photon distribution function. However, if we only allow Compton scattering the Kompaneets equation becomes relatively simple:

$$td\eta = n_e \sigma_T c \left(\frac{kT_e}{m_e c^2}\right) \frac{1}{x^2}\frac{\partial}{\partial x}\left[x^4\left(\frac{\partial \eta}{\partial x} + \eta + \eta^2\right)\right] \tag{4.37}$$

By making a change to a new independent variable, which is precisely the Compton y-parameter:

$$y \equiv \int_t^{t_0} \frac{k(T_e - T_\gamma)}{m_e c^2} n_e \sigma_T c dt \tag{4.38}$$

the low-frequency limit of equation (4.37) becomes a diffusion equation which describes the diffusion of photons from lower energies to higher ones via (inverse) Compton scattering:

$$\frac{\partial \eta}{\partial y} = \frac{1}{x^2}\frac{\partial}{\partial x}\left(x^4 \frac{\partial \eta}{\partial x}\right) \tag{4.39}$$

valid for $x \ll 1$. For linear deviations from a Planckian spectrum, $\Delta\eta$, we have [118]:

$$\frac{\Delta \eta}{\eta_0} = yx\frac{e^x}{e^x - 1}\left[x\frac{e^x + 1}{e^x - 1} - 4\right] \tag{4.40}$$

where $\eta_0 = (e^x - 1)^{-1}$ is the value of η at the same frequency, x. Unfortunately, the FIRAS data is at high frequency so that this solution is not good enough. Rather, numerical solutions must be used.

4.7.2 The Sunyaev-Zel'dovich effect for clusters

As mentioned earlier, inverse Compton scattering can occur whenever the CMB photons scatter off hot electrons ($T_e > 10^6 K$) giving the photons energy. This occurrence in the early universe was discussed above. However, in clusters of galaxies there is very often a component of hot electrons which will also induce spectral distortions in the CMB photons moving through the cluster. This has three effects: firstly the photons diffuse to higher temperatures in a way that depends both on the temperature of the electrons and the peculiar velocity of the cluster, and secondly, the photons are scattered, changing their direction. Finally, the increase in the temperature of the photons leads to second-order anisotropies on the sky. The second effect can be used to remotely sample the CMB at other places in the universe and hence we may begin to test the homogeneity of the universe. Further it can be used to make an independent estimate of the Hubble constant.

The Sunyaev-Zel'dovich (SZ) effect for clusters is usually divided between the thermal SZ effect and the kinematic SZ effect. The first reflects the thermal nature of the anisotropies due to the intrinsic temperature of the free electrons in the cluster core. The kinematic SZ effect is a result of the peculiar velocity of the cluster as a whole.

If one has the temperature of the electron gas in the cluster, through e.g. consistent X-ray observations, one can estimate the contribution to the SZ anisotropy from the peculiar velocity of the cluster, and hence gain an independent check on the velocity field estimates from e.g. POTENT. This however, is fraught with technical problems due to noise and is not a real prospect at this stage.

4.7.3 FIRAS and the future

FIRAS has given the following upper limits on spectral distortions from blackbody (through a simultaneous least-squares fit):

$$|y| < 2.5 \times 10^{-5} \tag{4.41}$$
$$|\mu_0| < 3.3 \times 10^{-4} \tag{4.42}$$
$$|Y_{ff}| < 1.9 \times 10^{-5} \tag{4.43}$$
$$\tag{4.44}$$

at the 95% confidence level. From equation (4.32) we see that this results in a limit to energy input of $\Delta E/E < 2 \times 10^{-4}$ for redshifts $10^3 < z < 8 \times 10^6$. As a result of FIRAS we now know that structure did not form through explosions and that the intergalactic medium doesn't have a uniform hot ($T \simeq 10^7 K$) component which had been previously suggested as a way of collisionally ionizing the intergalactic medium (IGM) to explain the Gunn-Peterson effect (the lack of distributed neutral hydrogen in the IGM). Further the 35 keV electrons producing the diffuse X-ray background must have a volume filling factor less than 10^{-4} [115]. This is because, for roughly constant electron density and electron temperature, the parameter y is proportional to the free electron column density which in turn can be related to the volume filling factor once a distribution model is specified.

However, perhaps the greatest thermal question remains unanswered: what was the global thermal history of the universe ? Was there recent reionisation or is the standard recombination scenario correct ? Further we need precise measurements of distortions from hot clusters to give peculiar velocities and to test the almost homogeneity of the universe (see later section on testing homogeneity).

To answer the first of these questions requires precise imaging of the spectrum at frequencies below 80 GHz which is the lower limit of the FIRAS experiment. Ground-based experiments probe the region down to frequencies of 0.3 Ghz and suggest that the mean temperature there is cooler than the $T_\gamma = 2.73 \pm 0.01K$ found by FIRAS, implying significant chemical potential distortions at these frequencies. However very little weight can be placed on these experimental results due to the huge error bars (relative to FIRAS). Thus in future we need accurate (i.e. better than or equal to FIRAS) measurements of the spectrum at long (centimeter) wavelengths.

Further, we need to analyse the angular variations of the spectrum. This is already possible with FIRAS since each parameter was calculated for every pixel. However, since they are essentially upper limits, the FIRAS data are not really suitable for e.g. a spherical harmonic approach, which would be able to put further constraints on the primordial power spectrum. Even at this early stage it is possible to say that $n < 1.6$, the spectral index in the primordial power spectrum. $n > 1.6$ would imply energy input via dissipation from matter acoustic waves which would cause an excessive chemical potential distortion not discussed above [74]. The prospect of large wavelength accurate measurements and definitive distortion angular spectra, make this field, and the field of CMB polarisation, a very promising one.

4.8 Polarisation

Up until now (barring the discussion on null cone formulations of anisotropy in section 4.3.4) we have characterised the radiation coming from the CMB by its intensity $I(\theta, \phi)$ on the celestial sphere. However, due to the anisotropy of scattering during the non-instantaneous decoupling phase, the CMB light is also partially polarised at a level that is typically much smaller (at the level of $1 - 10\%$) than the levels of temperature anisotropy. However, polarisation could provide very useful additional information that could be used to distinguish between the various cosmological models available. So how does one extend the standard formalism to include polarisation ?

4.8.1 An introduction to polarisation

Consider quasi-monochromatic light of average frequency ν moving along the z-axis. Then the components of the electric field are:

$$E_j(t) = a_j(t)\ \exp\left[i(\overline{\phi}_j(t) - 2\pi\nu t + 2\pi z/\Lambda)\right] \tag{4.45}$$

where $j = x, y$ and $\overline{\phi}(t)$ is the nearly constant phase of the wave. In the case of exactly monochromatic light, the amplitudes a_j and phases ϕ_j are all constant.

Now the rotation of the tip of E at constant z is determined by the difference:

$$\kappa = \overline{\phi}_x - \overline{\phi}_y \tag{4.46}$$

such that $sin(\kappa) > 0$ if the light is right-handed. By this one means that $\mathbf{E}$ appears to rotate clockwise to the receiver.

Now let us proceed to an operational definition of the Stokes parameters which will be used to completely characterise our polarised light. This way of defining them is one of the clearest I have seen and follows Born and Wolf 1959 [119]. Consider an E_y-compensator and a linear polariser. The first instrument will subject E_y to a phase retardation of ϵ radians relative to E_x. The linear polariser projects out the component of $\mathbf{E}$ at an angle θ counterclockwise to the x-axis. After passing through these two instruments the component of $\mathbf{E}$ in the direction θ is:

$$E(t,\theta,\epsilon) = E_x cos\theta + E_y e^{i\epsilon} sin\theta \tag{4.47}$$

with the observable intensity the following time average:

$$I_{trans}(\theta,\epsilon) = \langle E(t,\theta,\epsilon)E^*(t,\theta,\epsilon)\rangle_{t\gg\nu^{-1}} \tag{4.48}$$

We then perform 6 experiments to define the four Stokes parameters:

$$\begin{aligned}
\epsilon = 0 \quad &, \quad \theta = 0 \quad \rightarrow I(0,0) \equiv I_0 && (4.49)\\
\epsilon = 0 \quad &, \quad \theta = \pi/4 \quad \rightarrow I(\pi/4,0) \equiv I_{\pi/4} && (4.50)\\
\epsilon = 0 \quad &, \quad \theta = \pi/2 \quad \rightarrow I(\pi/2,0) \equiv I_{\pi/2} && (4.51)\\
\epsilon = 0 \quad &, \quad \theta = 3\pi/4 \quad \rightarrow I(3\pi/4,0) \equiv I_{3\pi/4} && (4.52)\\
\epsilon = \pi/2 \quad &, \quad \theta = \pi/4 \quad \rightarrow I(\pi/4,\pi/2) \equiv I_{right} && (4.53)\\
\epsilon = \pi/2 \quad &, \quad \theta = 3\pi/4 \quad \rightarrow I(3\pi/4,\pi/2) \equiv I_{left} && (4.54)
\end{aligned}$$

The results of experiments 5 and 6 are to produce right and left circularly polarised light respectively.

The Stokes parameters of the incident wave are then defined to be:

$$\begin{aligned}
I &= I_0 + I_{\pi/2} && (4.55)\\
Q &= I_0 - I_{\pi/2} && (4.56)\\
U &= I_{\pi/4} - I_{3\pi/4} && (4.57)\\
V &= I_{right} - I_{left} && (4.58)
\end{aligned}$$

or in terms of the amplitude and relative phases:

$$\begin{aligned}
I &= \langle a_x^2\rangle + \langle a_y^2\rangle && (4.59)\\
Q &= \langle a_x^2\rangle - \langle a_y^2\rangle && (4.60)\\
U &= 2\langle a_x a_y cos(\kappa)\rangle && (4.61)\\
V &= 2\langle a_x a_y sin(\kappa)\rangle && (4.62)
\end{aligned}$$

4.8.2 Polarisation of the CMB

In general polarisation can arise from several sources. The *vector* transport equation for the Stokes vector $\mathbf{I} = (I,Q,U,V)$ is:

$$\frac{d\mathbf{I}}{dz} = -\mathbf{K}(\mathbf{I}-\mathbf{S}) \tag{4.63}$$

where K is the absorpsion matrix which will usually depend on any magnetic fields present and S is the so-called source vector. Two broad areas in which polarisation is produced are when magnetic fields are present and when there is scattering from an anisotropic radiation field.

Now it is highly likely that a magnetic field existed at decoupling. We know that there is an intergalactic magnetic field from the Faraday effect it causes. Such a primordial magnetic field, if non-negligible, has significant implications. Not only do almost all present CMB calculations exclude magnetic fields, but due to the vector nature of the field, it introduces a local anisotropy which is not amenable to standard isotropic/homogeneous treatments, unless one assumes that the magnetic fields in neighbouring domains are not correlated and one considers averaging scales much larger than these domains.

In this section we will not attempt to look at the polarisation that might result from such early magnetic fields, leaving that to section (4.8.4). Rather, we will examine polarisation due to the small anisotropies in the radiation field at decoupling resulting from the primordial perturbations in the energy density. Of this anisotropy, the quadrupole is by far the most important multipole moment.

Now to first order one may regard recombination as instantaneous. This is further justified by the "compensation effect": radiation cools at the same rate when decoupled from matter as when it is coupled to matter, so that neglecting the extra distance traveled by some photons relative to others makes no difference in their observed temperature at first order.

However in truth the surface of last scattering (which to fairly good approximation coincides with the surface of recombination) has a thickness of about $\Delta z \simeq 80$ if $z_{SLS} \simeq 1000$. Photons scatter from the free electrons primarily through Thompson scattering with the cross-section: [120]

$$\sigma_T \propto |\epsilon \cdot \epsilon'|^2 \tag{4.64}$$

where ϵ, ϵ' are the initial and final photon *vector* polarisations respectively.

At this stage we must actually calculate quantities using the *vector* Boltzmann equation for the four Stokes parameters introduced earlier. However, because we are not treating the case of a magnetic field, the polarisation induced will be linear (i.e. due to anisotropic scattering so there is no circular polarisation, eq. 4.58). Hence we will not need the last Stokes parameter, V. In the case of scalar (i.e. pure density) perturbations we do not require U either [121]. For tensor perturbations this is not true, but one can reduce the coupled system from three to two dimensions by a change of variables [120]. It is at this stage that the literature splits into a myriad of different notations and variables which are often physically unclear. As most of the results quoted are numerical solutions they do not provide us with much physical insight.

We can, however, gain some physical insight using an extension of the perturbative expansion of the Boltzmann equation hierarchy (i.e. after Fourier and harmonic-ℓ expansions), pioneered by Hu and Sugiyama [104]. This expansion is done in terms of the inverse optical depth, κ^{-1}, which corresponds to the mean time, τ_C, between scatterings. Including polarisation we may then summarise the situation qualitatively as follows [122].

At zero order in τ_C, we have perfect (tight) coupling between the photons and electrons and hence the photon distribution is isotropic in the electron rest frame and there is no polarisation because there is no temperature anisotropy. At most if one is moving in another frame, one can see a dipole due to the relative velocities.

At first order in τ_C one finds a quadrupole component of the polarisation that is proportional to the quadrupole of the temperature anisotropy. The temperature quadrupole itself is proportional to the temperature dipole. The temperature moments for $\ell > 3$ are essentially zero in this case at decoupling. Power is then transferred to the higher multipoles during free-streaming due to couplings in the Boltzmann hierarchy.

4.8.3 Temperature - Polarisation correlations

When examining only the temperature anisotropy maps of the CMB, one is faced with the question: is it possible to determine all relevant parameters from the temperature anisotropies alone ? Even without contamination the parameter space turns out to be degenerate and one has to resort to statistical maximum-likelihood methods. However, polarisation maps when combined with the anisotropy maps, allow one to extract more information. It turns out, that depending on the type of perturbation (scalar or tensor) responsible for the CMB anisotropy, the polarisation vector field will behave differently in the neighbourhood of hot and cold spots. For scalar perturbations, hotspots are surrounded by a vector field tangent to circles of constant temperature, while the vector field is radial about cold spots [120]. For tensor perturbations the exact opposite occurs (assuming a flat, $\Lambda = 0$ background). This is shown in figures (4.8.3) and (4.8.3) for square 20° maps of the sky with the correlated component of the polarisation field overlaid.

In general the polarisation consists of a part which is correlated with the anisotropy, Q_c, and a part which is not, Q_U. In the case of tensor perturbations, the correlated part constitutes a much larger fraction of the total polarisation than in the scalar case. The correlated part is used in the two maps shown.

A statistical quantification of these facts is the cross-correlation, $\langle Q\, T\rangle$, where Q is the Stokes parameter defined in the previous section and T the temperature. The above geometric difference between the polarisation-temperature fields for scalar and tensor perturbations is reflected in the fact that the cross correlation function has roughly opposite sign in the cases of scalar and tensor perturbations (see figure 4.8.3). This method does in principle offer a way of determining whether tensor or scalar perturbations were responsible for the anisotropy but because of the very small signal, such experiments will be very noisy and difficult. Further, there is a confusion due to Ω that has not been included. As discussed in earlier sections, because the ISW effect contributes significantly to the anisotropy when $\Omega < 1$, and is produced at low-redshifts, there will be a significant drop in the size of the cross-correlations discussed in this section (since no polarisation is produced at low redshifts in standard reionisation scenarios). Further, since the SW and ISW effects occur with opposite sign, hot spots and cold spots will be approximately interchanged. This inversion will be a significant problem for the test proposed by Crittenden *et al* [120], discussed here, but could still be very useful once Ω is known and high-resolution all-sky maps have been obtained for both temperature and polarisation.

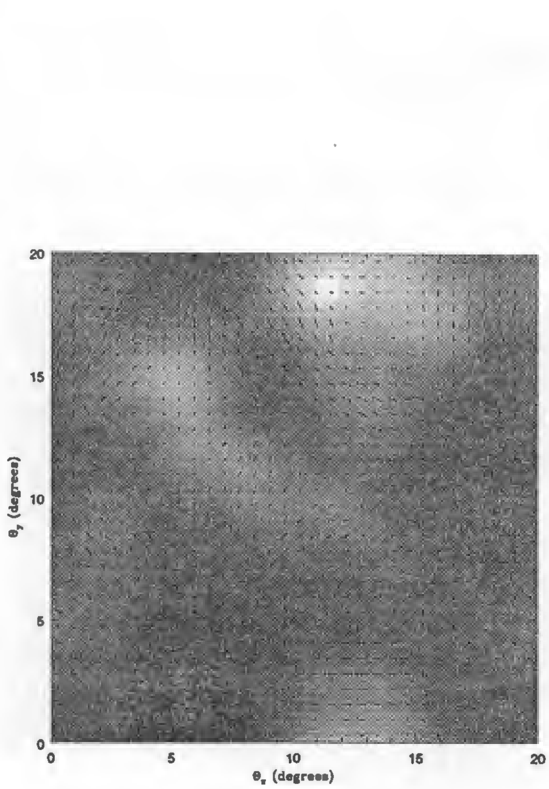

Figure 4.5: A $20° \times 20°$ map of the anisotropy from adiabatic scalar perturbations with correlated polarisation field. Note that the lighter shading indicates hot spots. The polarisation field is radial around hot spots and tangential about cold spots. From Crittenden *et al* (1995) [120].

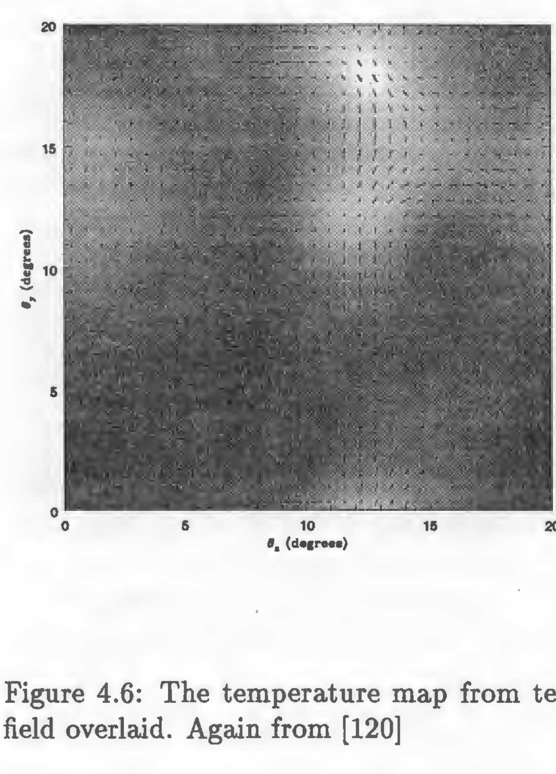

Figure 4.6: The temperature map from tensor perturbations with correlated polarisation field overlaid. Again from [120]

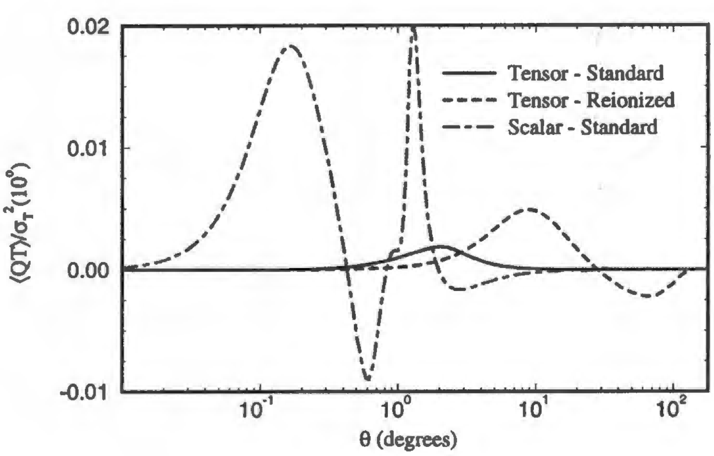

Figure 4.7: The QT cross correlation on a $\phi = 0$ slice. Shown are the scalar polarisation which is dominant at small angles, and the tensor polarisation in standard and reionised scenarios. Reionisation significantly increases the polarisation and the scalar, tensor cross correlations are seen to have opposite sign on the scale of a few degrees. For angles larger than 30° only the reionised tensor cross correlation remains, and even then at less than 0.5% of the variance $\sigma_T^2(10°)$ of the temperature anisotropy [120].

4.8.4 Faraday rotation of the polarisation vector by primordial magnetic fields

The existence of magnetic fields of order $3\mu G$ in galaxy clusters and spiral galaxies is well established, but a long-standing mystery [123]. There are two extreme ideas: either the fields were generated primordially or they were produced by exponential dynamo amplification of a very small seed field. In the first case they could have rather dramatic implications for structure formation because of mode coupling between density, vorticity and the magnetic field. it also introduces theoretical problems since magnetic fields are vectors and hence a coherent magnetic field on the scale of the horizon would destroy isotropy, making one of the Bianchi models a better approximation than the traditional FLRW models. Thus the question arises: "How could one detect such a coherent magnetic field of primordial origin ?"

As we saw from eq. (4.63), any magnetic fields appear directly in the absorpsion matrix **K**. One way we might detect a coherent magnetic field that was present at last scattering is to search for changes in the Faraday rotation of the polarisation vector of the CMB light across the sky. Obviously this is a third generation CMB experiment - first was the discovery of anisotropies in the CMB (1992). Second we have to obtain positive detections of the polarisation of the CMB which are at levels of $1-10\%$ of the anisotropy (circa 1997-2002 ?) and then thirdly we can look for polarisation maps and coherent rotations of the polarisation vector (circa 2002-2008 ?). Since this is a very small effect in general it will be very sensitive to contamination. Fortunately, however, gravitational lensing leaves the polarisation vector invariant [142].

This problem has recently been addressed [106]. They present a calculation which is an extension of the formalism presented in section 4.3.4. However we can avoid detailed calculations and estimate roughly what the approximate size of the effect will be.

First we note that there are two ways that one can extract information about magnetic fields from polarisation maps: firstly one may look at polarisation on different angular scales and secondly one can note that the magnitude of the rotation is frequency dependent, so that multi-frequency observations will yield the required magnetic field.

Now consider monochromatic radiation of frequency ν travelling in the direction of the unit vector **q**, through a plasma with magnetic field **B**. The total rotation in the direction of the (linear) polarisation vector will be:

$$\varphi = \int_E^R \frac{e^3 x_e n_e}{2\pi m^2 \nu^2}(\mathrm{B}\cdot\mathrm{q})dt \tag{4.65}$$

where e, m, n_e, x_e are the electron charge, mass, number density and ionisation fraction, respectively. Now the magnetic field behaves like vorticity and decreases like $(1+z)^2$, so that B/ν^2 is a constant. Further, since a good definition of the surface of last scattering is that it is the surface of unit optical depth due to Thompson scattering, we may write $\int x_e n_e dt = 1/\sigma_T$ (see eq. 4.12). By averaging $\varphi^2 \propto B\cos^2(\theta)$, θ the angle between **B** and **q**, over all possible orientations of **B** one obtains the *rms* rotation angle [106]:

$$\langle\varphi^2\rangle^{1/2} \simeq \frac{e^3 B_0}{2\sqrt{2}\pi m^2 \sigma_T \nu_0^2} \simeq 1.6^\circ \left(\frac{B_0}{10^{-9}\ \mathrm{Gauss}}\right)\left(\frac{30\ \mathrm{Ghz}}{\nu_0}\right)^2 . \tag{4.66}$$

where the subscript 0 denotes evaluation today. An amazing feature of this result is that it doesn't depend on cosmological parameters as long as it remains true that B/ν^2 is constant.

Now looking at Faraday rotation of light from quasars, it is possible to constrain a coherent magnetic field on horizon scales. This has been done and the upper limit $10^{-9}G \times (\Omega_{IGM}h/0.01)^{-1}$ has been found [124], where as usual h is the Hubble constant in units of $100 km/s/Mpc$. Therefore if we put $B_0 = 10^{-9}$ G we find a *rms* rotation of about $1.6^{o}cm^{-2}$ at 30 GHz. This is a very simple calculation which contains no details about the physics of last scattering, which could change the answer by a non-negligible amount and is worth investigating in detail. The ν^{-2} variation in frequency is sharp enough to enable one to consider detection of this effect within a decade, if magnetic fields were mainly primordial in origin, thus potentially solving one of the long-standing mysteries of galaxy formation.

4.9 Is the Universe Almost Homogeneous ?

As a final closing topic to this chapter it is appropriate to discuss the impact the CMB will have on testing the Copernican principle, which is perhaps the crucial hurdle in the way of finally converting cosmology into a science.

As mentioned briefly earlier, there is a strong connection between the almost-Copernican principle and the assumptions of Gaussianity, ergodicity and statistical isotropy and homogeneity. One of these is required to be able to make deductions about cosmological information. Statistical isotropy and homogeneity could almost be taken as a definition of the almost-Copernican principle.

There is an interesting subtlety however. Assuming such an almost-Copernican principle is not sufficient within a statistical context. Ergodicity is required to make useful statistical deductions about the universe from our limited vantage point. However, as noted in the section on ergodicity, this is impossible on the celestial sphere since it has finite volume, *even if* the perturbations are statistically isotropic and spatially homogeneous. In this sense, ergodicity is a stronger requirement than an almost-Copernican principle. The difference between the two is encoded in the "fudge factor" - cosmic variance.

The question then arises as to whether it is possible yet, to prove that the universe is almost-FLRW, i.e. to show almost-isometry about many points in the given spacetime region that we want to show has almost-FLRW metric and dynamics. Let us formalise this by considering an open region Σ about our worldline which is the region that we can effectively probe with observations of large-scale structure, say $z < 5$. The key to proving almost-homogeneity lies in examining the isotropy or anisotropy of the universe at other spacetime points. In particular, if we could see the CMB at other points we would be in a strong situation, since we could then invoke the almost-EGS theorem.

The Sunyaev-Zel'dovich effect provides exactly such a telescope, albeit rather coarse. The idea is simple: inverse-Compton scattering involves an angular scattering of the CMB photons. This distribution of scattered photons gives us a shadowy look at the CMB emitted from worldlines that we could not otherwise probe. Now if the CMB at the cluster where the scattering occurs, is highly anisotropic with $\Delta T/T \sim 1$, say, then the monopole coming through the cluster will differ significantly (after accounting for redshift differences) from the 2.726 K we see. Thus the CMB at least offers a possibility of testing the homogeneity of the universe on the largest scales.

We now move on to issues of gravitational lensing.

Chapter 5

A Bird's-Eye view of Gravitational Lensing

"Things derive their being and nature by
Mutual dependence and are nothing
In themselves"

Nagarjuna

5.1 Introduction

Our aim in this chapter is to briefly review the current situation in gravitational lensing, particularly with respect to cosmological distance measures, and hence to provide the background for the following two chapters.

The study of distances in inhomogeneous universes is very difficult and lies near the boundaries of applicability of many techniques of General Relativity and cosmology. On the one hand one has the simple FLRW distance relations which exclude lensing. One then has a hierarchy of more complex exact solutions and approximations which quickly become more intractable (see figures 5.1, 5.2). On the other hand one has numerical approaches which again require approximations and are limited by computing power constraints.

Gravitational lensing is a new field but there are many approaches to the problem all of which have different advantages and disadvantages. In much of this chapter we will focus on attempts to find accurate distance measures which include the effects of gravitational lensing. In general, all one can safely assume is that the geometric optics limit holds almost all of the time, implying that photon number is conserved [142].

However, a powerful tool exists in attacking the problem of observations in a general universe: the optical scalar equations.

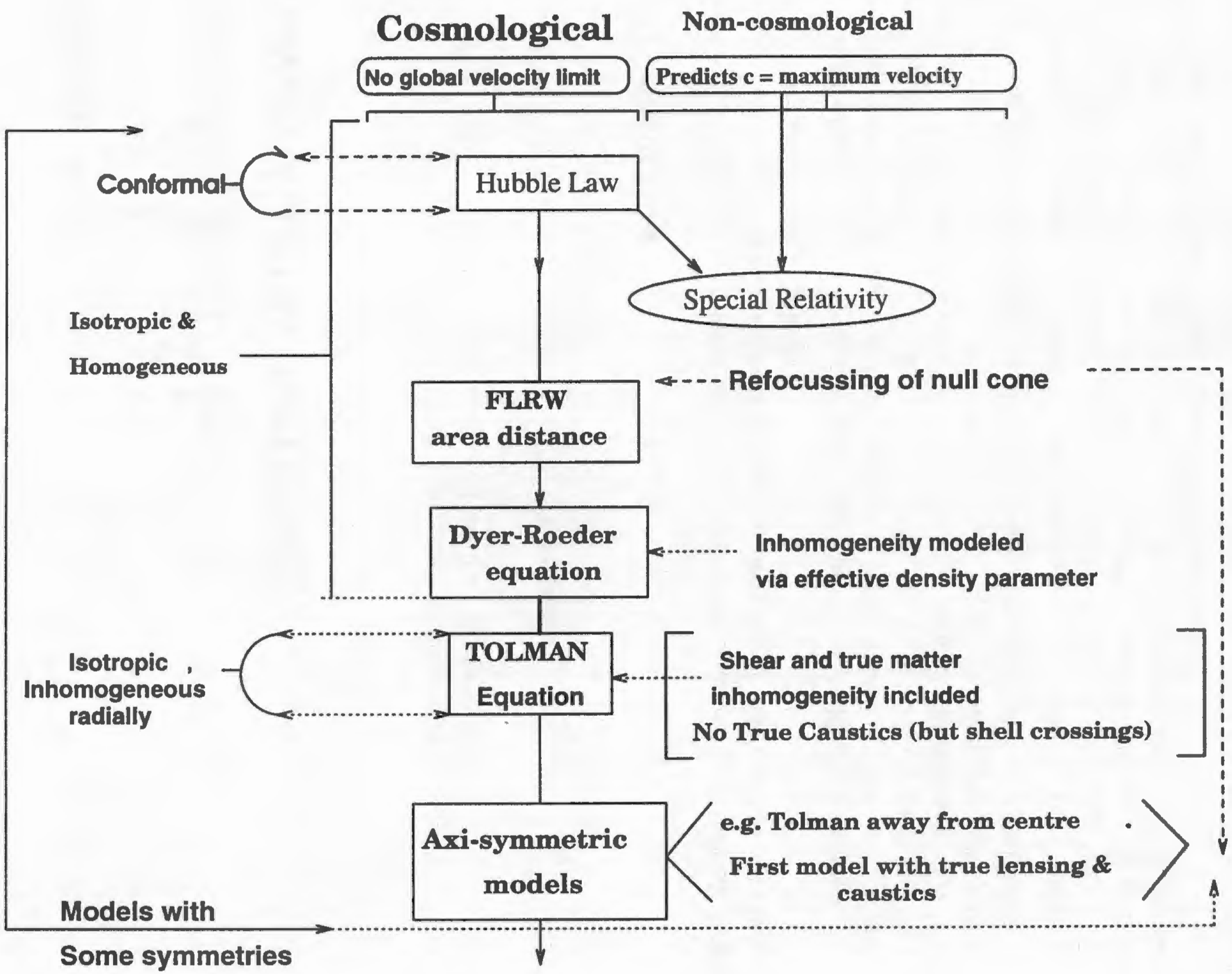

Figure 5.1: The hierarchy of area distances in models with some symmetry.

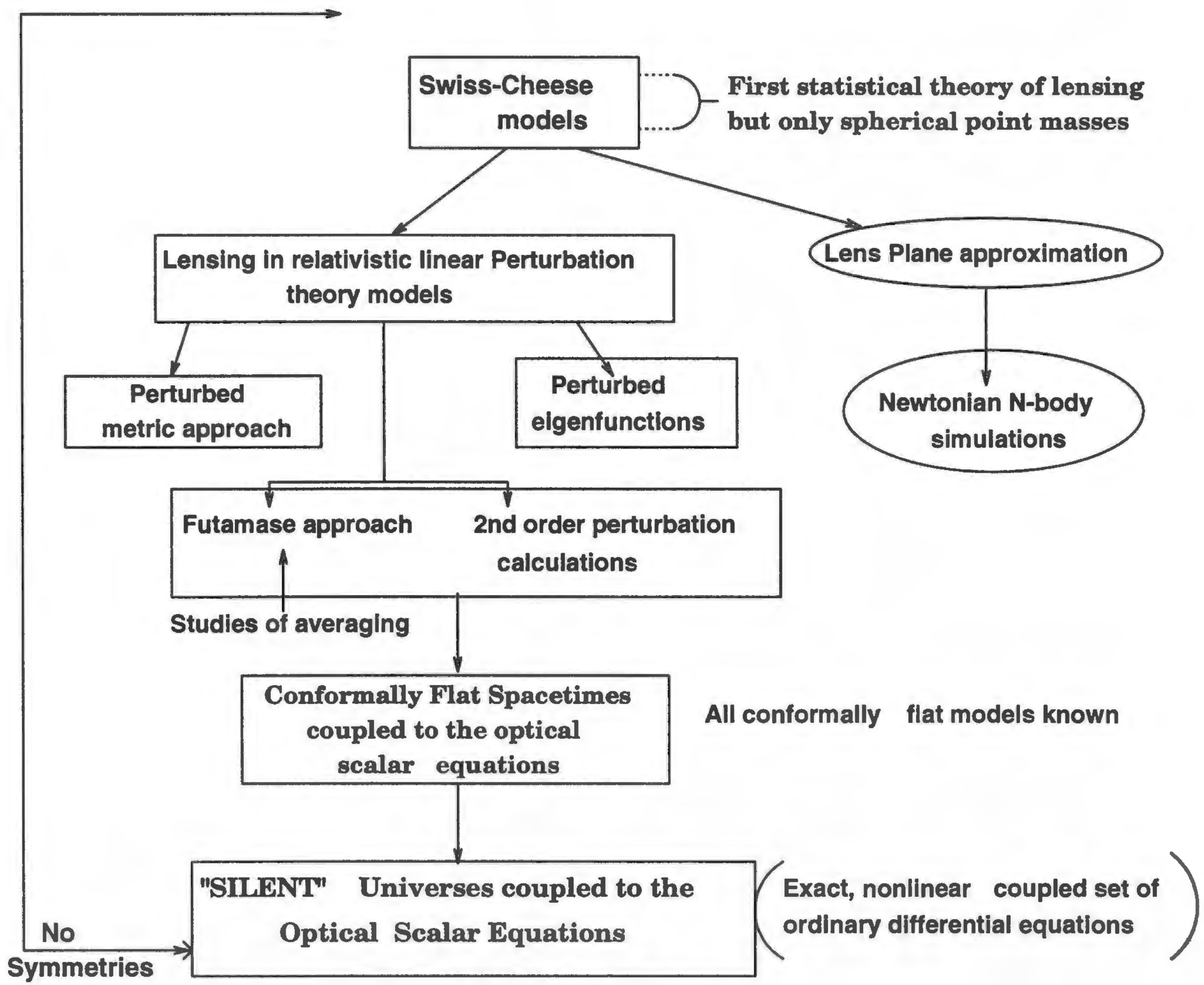

Figure 5.2: The hierarchy continued from figure 5.1. The models pocess no symmetries, and are hence much more difficult to work with.

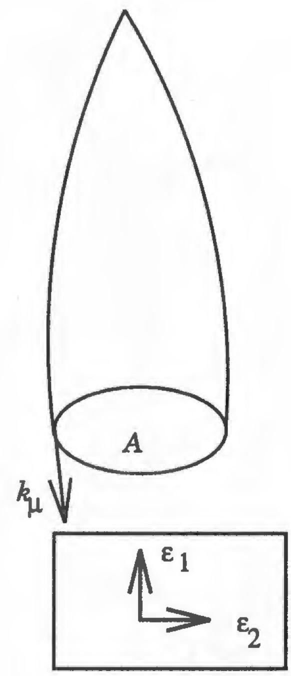

Figure 5.3: A schematic diagram of a ray bundle of area A, wave vector k_μ, and complex screen vectors ϵ_μ.

5.1.1 The optical scalar equations

In general one is interested in the behaviour of infinitesimal bundles of null geodesics, which leads to the definition of the area distance, r, as the square root of the area, A, of the ray bundle; the intensity of the light varying inversely with bundle area, see fig. (5.3).

The restriction to infinitesimal bundles allows the estimated area distance to vary in different directions, which in an inhomogeneous universe, is what generically occurs. The area distance relates the physical size (say of an inhomogeneity on the surface of last scattering), l, at a given redshift, z, to the angle, θ, that it subtends in the sky (assuming the bundle has not passed through conjugate points [1]:

$$l(z) = r\theta \tag{5.1}$$

We can relate r to the optical quantities of the bundle: the expansion, θ, the (complex) shear, σ, and the vorticity, ω, through

$$\frac{1}{r}\frac{dr}{d\nu} = \theta \tag{5.2}$$

with ν an affine parameter and:

$$\frac{1}{2}\sigma_{\mu\nu}\sigma^{\mu\nu} = |\sigma|^2 \qquad \theta = \frac{1}{2}k^{\mu}{}_{;\mu} \qquad \omega^2 = \frac{1}{2}k_{(\mu;\nu)}k^{\mu;\nu}\ ,$$

with $\pm|\sigma|$ the eigenvalues of the shear tensor and ; denoting covariant derivative.

[1]Conjugate points occur when distinct null geodesics intersect each other. They are the birth places of caustics.

The evolution of these optical scalars is given by the nonlinear system [155, 142]:

$$\dot{\theta} + \theta^2 - \omega^2 + |\sigma|^2 = \frac{1}{2} R_{\mu\nu} k^\mu k^\nu \tag{5.3}$$

$$\dot{\omega} + 2\omega\theta = 0 \tag{5.4}$$

$$\dot{\sigma} + 2\theta\sigma = \frac{1}{2} C_{\mu\nu\alpha\beta} \epsilon^{*\mu} \epsilon^{*\alpha} k^\nu k^\beta \tag{5.5}$$

where $C_{\mu\nu\alpha\beta}$ is the trace-free Weyl tensor responsible for tidal distortion. k_μ and $\epsilon^{*\nu}$ are respectively the wave vector, and a complex null vector orthogonal to the radiation (the "screen vectors".

In cases where there is a symmetry, these equations simplify greatly, and in the case of FLRW and Dyer-Roeder models, become a single second order equation for the area distance with redshift, the shear being zero. In fact, the shear of the ray bundle is zero in any conformally flat spacetime, since the initial conditions at the vertex of the null cone must include $\sigma \rightarrow 0$, and from equation (5.5) this implies that $\sigma = 0$ at all times. This means that the ray bundle's evolution is only determined by the $R_{\mu\nu} k^\mu k^\nu$ term which is known as the Ricci focusing term since it describes the effect of the energy density interior to the ray bundle. The Weyl term is the source for shear and it is this shear which causes the ray bundle to become elliptical due to causal non-local tidal forces. It is easy to show by combining eqs. (5.2) & (5.3) [142] that only $|\sigma|^2$ appears in the evolution equation for the area distance, implying that addition of any shear increases the convergence (decrease of area) of the ray bundle prior to the formation of conjugate points. Thus converting a homogeneous matter distribution into an inhomogeneous one through clustering has two effects: the Ricci focusing effect changes because the mean energy density inside the beam changes and second, the matter density peaks cause tidal shearing of the bundle. For a bundle propagating between galaxies, the two will have opposite effects on the area of the bundle.

In the geometric optics limit, the vorticity of the ray bundle is zero, $\omega = 0$. In the case where the shear or Ricci focusing is sufficiently strong, the ray bundle can be forced to converge locally. In this case the area goes to zero, the geodesics intersect and hence the luminosity diverges formally at these conjugate points. After this the bundle and the wavefront becomes multiply sheeted and there are multiple images of the source. This is the realm of strong lensing.

5.2 Strong Lensing

Because of the difficulty of integrating the geodesic or optical scalar equations in realistic cases (where the Weyl tensor may not even be known) and particularly because of the problem of extending the integration analytically through conjugate points, a simplified model has gained extensive use; the so-called lens approximation.

In the lens plane approximation, it is assumed that the lensing mass has a spatial extension well approximated by a $2-d$ plane. On this plane there is a parameter, called the critical surface mass density, Σ_{crit}, which characterises the surface density on the plane needed to produce multiple images for a source and observer at given distances from the lens plane. It is defined by:

$$\Sigma_{crit} = \frac{c^2 D_s}{4\pi D_d D_{ds}} \tag{5.6}$$

where D_s, D_d, D_{ds} [2] are respectively the area distance from the observer to the source, lens (deflector) and the area distance from the lens to the source, with the correct null cone normalisations at the source. In an Einstein-de Sitter universe, the area distance is:

$$D(z) = \frac{2}{(1+z)}\left[1 - \frac{1}{\sqrt{1+z}}\right] \tag{5.7}$$

This gives us $D_s \equiv D(z_s)$ and $D_d \equiv D(z_d)$ in this model. D_{ds} is given by [142]:

$$D(z_d, z_s) = \frac{2}{(1+z_s)}\left[\frac{1}{\sqrt{1+z_d}} - \frac{1}{\sqrt{1+z_d}}\right] \tag{5.8}$$

in the EdS model. This gives the critical surface mass density:

$$\Sigma_{crit} = \frac{c(1+z_s)}{2\pi}\left[\frac{(1+z_d)^2}{(1+z_s)^2}\right]\frac{[1+z_s-\sqrt{1+z_s}]}{[1+z_d-\sqrt{1+z_d}]}\left[\frac{1}{\sqrt{1+z_d}} - \frac{1}{\sqrt{1+z_s}}\right]^{-1} \tag{5.9}$$

Σ_{crit} is plotted in fig. (5.5, (5.6) for constant lens redshift and source redshift respectively. As can be seen, as the source redshift is increased, the critical surface mass density decreases rapidly and then flattens off asymptotically. For astrophysical situations this gives a good indication of whether or not caustics will form. However, in the cosmological arena this is probably not very good because it ignores the cumulative effect of the matter structures traversed before reaching the lens plane [217], which gives the incoming ray bundle area a large variance about the mean of a suitable average and therefore makes Σ_{crit} a crude statistic that generically underestimates the probability of forming multiple images. It is these facts which make it plausible that caustics possibly play an important role in the CMB on some angular scales, and is the main focus of the next two chapters.

5.2.1 Caustics and catastrophes

As discussed earlier, the lens plane approximation reduces the Einstein field equations to the study of smooth mappings from $\mathbf{R}^2 \to \mathbf{R}^2$, from the source plane to the lens plane. The study of such maps and their singularities (caustics) is the subject of catastrophe theory [142]. The generic caustic for lensing studies is the swallowtail catastrophe which can be identified locally with the projection of the past null cone into the spatial slices in the neighbourhood of the event of formation of the caustic (see figure 5.7) [125].

Caustics induce non-differentiable subsets into the wavefront because of their sharp edges where, at least formally, the area of a thin ray-bundle goes to zero, giving infinite intensity in the point-source, geometric optics limit. In practise, finite source size and coherence effects from the wave-nature of light render the intensity large but finite, at all caustic points. We refer the reader to e.g. [142] for more details on caustics and leave till the next chapter an assessment of their importance to high-redshift observational cosmology.

[2] We use the notation of Schneider *et al* [142].

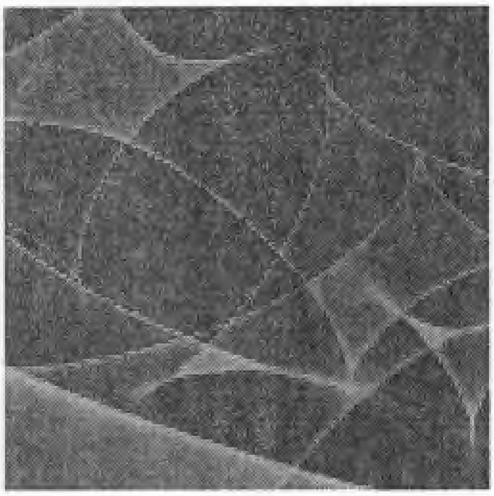

Figure 5.4: The critical curves from a typical strong-lensing ray-shooting simulation. The number of images of a discrete source changes by 2 every time one crosses a caustic curve. From the Max-Planck web site; http://www.mpa-garching.mpg.de/Lenses/GRLens.html.

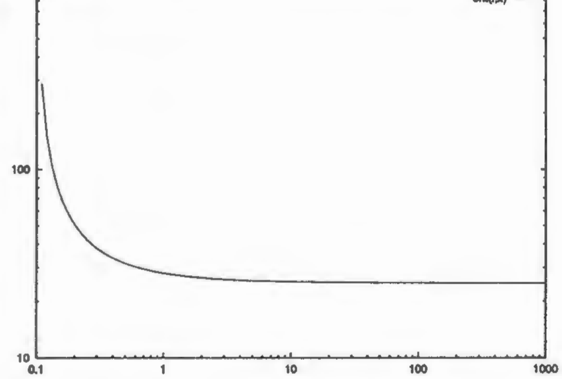

Figure 5.5: Σ_{crit} as a function of source redshift for fixed lens redshift ($z = 0.1$) in the Einstein-de Sitter model. The surface mass density needed to produce multiple images of the CMB is much less than for a low-redshift source, even for a single lensing object. The total probability for multiple lensing should be integrated over the total matter distribution.

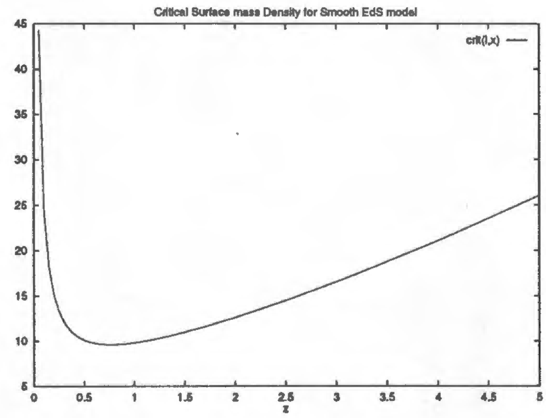

Figure 5.6: Σ_{crit} as a function of lens redshift for the sources in the CMB, assuming that $z_{SLS} = 1000$, i.e. no reionisation.

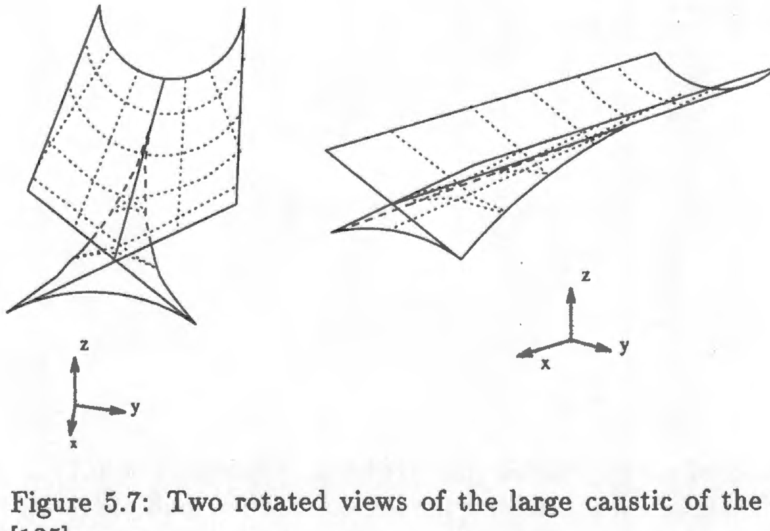

Figure 5.7: Two rotated views of the large caustic of the swallowtail catastrophe. From [125].

5.3 The hierarchy of area distances - models with symmetries

Here we give a necessarily concise review of current understandings of area distances in cosmology. We will not discuss the Special Relativistic case, but simply note that it is completely wrong to use it when discussing astrophysics in an expanding universe. This comment is not without point since it is quite common for astronomers to say things such as "a quasar with redshift $z = 4.5$ is moving away from us at 98% the speed of light", while in truth, a model for the expanding universe gives its proper velocity as greater than 2c [209]. In General Relativity the speed of light is not a global velocity limit. Of course locally it is still the ultimate limit.

5.3.1 Isotropic and homogeneous spaces

The conformal Hubble law

The simplest area distance is simply the linear Hubble law. The linearity ensures that it is a conformal mapping from redshift to real space, so that angles between structures at small redshifts in redshift space are true reflections of the angles in real space. However, this is not true for General Relativistic laws, or when one includes the peculiar velocities of objects. The first deviations from the linear shape come in FLRW models from q_0, the deceleration parameter, the energy density Ω of the universe and the cosmological constant Λ. Thus if one knows Ω and Λ, and in addition has standard candles at reasonably large redshifts (such as Type Ia supernovae), one can apparently estimate q_0 accurately. It is known that weak gravitational lensing is not very important (error $\leq 10\%$) [218], while the effects of strong lensing were investigated in the Swiss-Cheese model and are expected to be important for supernovae at $z > 1$ [219].

The FLRW area distances

These are fully relativistic and as mentioned before exhibit no limit in the velocity-redshift diagram, i.e. $v \to \infty$ as $z \to \infty$. As a simple example the Einstein-de Sitter area distance (in units of the Hubble radius c/H_0 is:

$$D(z) = \frac{2}{(1+z)}\left[1 - \frac{1}{\sqrt{1+z}}\right] \tag{5.10}$$

In the case of hyperbolic spatial geometry ($k = -1$), the volume of a region of constant solid angle grows exponentially with redshift due to the exponential divergence of geodesics in this case. Thus the area distance corresponding to a given angle is much larger in an open $\Omega < 1$ universe than in the flat analogue. However, these do not include any effects of gravitational lensing and hence are not particularly interesting for us except for use in comparison with models with lensing.

The Dyer - Roeder approximation

A step towards including lensing can be taken while retaining the isotropic and homogeneous nature of the solution essentially by neglecting the shear on the ray bundle and assuming that the affine parameter is unchanged by lensing. The original derivation assumed that one can model the distribution of matter as made of clumps which constitute $(1-\alpha)\Omega$ of the total density parameter ($\alpha \leq 1$). These effects of the clumped matter are then neglected. The ray bundle is then effectively propagating in a universe of density parameter $\alpha\Omega$, while the dynamics of the (FLRW) universe is determined as usual by the full Ω. Because the Ricci focusing on the ray bundle (due to the matter contained within the bundle) is less, the area distance is greater than in the true FLRW case, an effect known as "shrinking" [130, 212].

This approximation is only valid in two extremely different cases:

(1) When the global character of the wavefront exhibits a mean behaviour that can be modeled by an effective density parameter and one is only interested in the coarsest average behaviour of the area distance.

(2) When one is considering a small ray bundle on astrophysical scales which propagates far from any clumps; i.e. through an essentially constant background density intergalactic medium, & the shear is negligible.

The second is the traditional approximation in lensing studies, while the first is emerging as an appropriate "coarse-grained" picture for treating the whole past null cone in the presence of a statistically isotropic distribution of caustics, as will be discussed in the next chapter. In this sense it is an "averaged", statistical equation for the mean area distance $\langle r \rangle$.

If we substitute the assumptions and approximations into the optical scalar equations (5.5), one can derive the Dyer-Roeder equation:

$$P(z)\ddot{r} + Q(z)\dot{r} + \left[\frac{3}{2}\Omega_d\alpha\right] r = 0 \,, \tag{5.11}$$

$$P(z) = (1+z)(1+\Omega_d z) \,, \tag{5.12}$$

$$Q(z) = (3 + \frac{\Omega_d}{2} + \frac{7\Omega_d z}{2}) \,. \tag{5.13}$$

The corresponding initial conditions are [143]:

$$r(z=0) = 0 \,, \qquad \dot{r}(z=0) = 1 \,. \tag{5.14}$$

which can be solved in many cases in terms of hypergeometric functions. However, since these have themselves to be generated numerically we will not discuss these solutions which are given in the next chapter.

5.3.2 Isotropic and axially symmetric (LRS) models

Obviously the Dyer-Roeder (DR) approximation is an improvement over the FLRW distances even if the successful application range is fairly limited. However, the DR result is essentially a 2-density FLRW solution and hence doesn't really describe the effects of inhomogeneity. The simplest generalisation of this is to consider spherically symmetric but radially inhomogeneous models. This is a highly non-trivial extension however, particularly where cosmological observations are concerned. It is discussed in detail in chapter (6).

The canonical cosmology in this class is the Lemaître-Tolman-Bondi (LTB) model with the observer at the central worldline. The best approach is perhaps not given by the optical scalar equations but rather by integrating the radial null geodesics. This has recently been made a great deal easier by the realisation that the "gauge-freedom" in the LTB model can be specified on the past null cone in such a way as to simplify the equations for the null geodesics [213]. This is done by making a single past null cone flat (conformal coordinates). This allows one to obtain several very interesting results, particularly regarding the non-trivial effects of angular and spatial averaging on the area distance, the main issue under investigation in chapter (6).

The beauty of the LTB model is that the area-distance is in fact the transverse scale factor, R. Thus the equations of motion for the metric give the equation for the evolution of the total (or average) wavefront area. The inhomogeneity is specified on the null cone, denoted by the function $\hat{\rho}$ and the area satisfies the equation [213]:

$$\hat{R}(1-\frac{d\hat{R}}{dr})\frac{d^2\hat{R}}{dr^2}-\frac{1}{2}\hat{R}(1-\frac{d\hat{R}}{dr})^2+4\pi\hat{\rho}\hat{R}^2=0 \tag{5.15}$$

where the $\hat{}$ indicates that the quantities are evaluated on the null cone. This is the Tolman analogue of the Dyer-Roeder equation. This equation is completely non-linear in contrast, however, encoding the non-linear matter shear. A further important feature is that it is not possible to cast this as a second-order differential equation in z since the redshift is itself related to the radial coordinate r in a non-linear way. Finally, one still has to worry about shell crossings and redshift disordering. The first occurs when successive shells of matter break through each other. Obviously this is not physical but is in fact the spherically symmetric analogue of the shell crossings in the formation of Zel'dovich "pancakes" in a pressure - free medium. Redshift disordering occurs when the mapping from $r \rightarrow z$ is not one-to-one.

We should note that the lensing in this case is purely radial - as inhomogeneity is introduced the null geodesics exhibit no angular fluctuations. Instead the past null cone develops a wavy geometry due to time delays induced by the matter inhomogeneities, so that the area distance may vary considerably from that in the FLRW models. Now there are no true caustics and multiple lensing for the central LTB observer. However, if we move away from the central worldline, the spherical symmetry is broken to an axial-symmetry. This is more general and is the first model we have considered which actually allows for angular fluctuations in the geodesics due to spatial inhomogeneity, and hence it is the first model with real caustics.

Little work in modern lensing theory has been done in these models since realistic calculations of lensing effects must account accurately for the observed large scale structure and CMB, which are explicitly statistical theories. Encoding these statistics in these non-linear cosmologies with minimal axial symmetries is very difficult and asymmetric linear perturbations theory has been substantially preferred.

5.4 The area distances in models without symmetry

5.4.1 Swiss-Cheese models

The Swiss-Cheese models are exact, nonlinear models obtained by e.g. taking a smooth FLRW background, removing spherical vacuoles and matching a spherically symmetric mass distribution to the inside of each vacuole with a Schwarschild metric. As a result they are fully anisotropic and inhomogeneous in general. It is possible to formulate the optical scalar equations in a reasonably tractable and realistic form for numerical study. The best analysis of this model to my knowledge was performed by Dyer & Oates (1988) [199].

The advantage of the Swiss-Cheese model is that the Weyl tensor is non-zero only inside the vacuoles, while the Ricci focusing term is non-zero only outside the holes. In both cases the contributions are known exactly. Dyer & Oates proceed by calculating the expected statistical values for the rate of shear which is obtained by summing over the Weyl tensor contribution from each vacuole the beam passes through. Further they assumed that the vacuoles were uniform randomly distributed, something we know to be invalid. Nevertheless, they were able under this simplifying assumption to calculate the step in the shear at each hole.

The optical scalar equations become third order equations for the area distance in terms of redshift. The results of [199] showed that caustics demagnify most sources in the sky while strongly magnifying a few rare sources. Although they did not study the case of all-sky averaging (due to its extreme difficulty in the Swiss-Cheese model), their work strongly supports the idea that caustics cause shrinking, as conjectured in chapter (7).

Linear perturbation theory

One of the advantages of the metric perturbation approach discussed in detail in chapter (4) on the CMB, is that it allows one to integrate the perturbed null geodesics in terms of deviation vectors in each spatial slice, (until conjugate point formation). The equation of motion of the wave vector tangent to the geodesic is given by [126]:

$$\frac{d\mathbf{n}}{dl} = 2\mathbf{n} \times (\mathbf{n} \times \nabla\phi) \tag{5.16}$$

where l is the comoving path length along the photon geodesic. This is valid for stationary, linear potential corresponding to the (gauge-dependent) linear overdensity δ via the Poisson equation:

$$\nabla^2\phi = \frac{3}{2}H^2\Omega S^2\delta \tag{5.17}$$

With the above evolution equation for $\mathbf{n}$, one can construct statistics about the difference in angular fluctuations between neighbouring geodesics which indirectly encodes information about the change to the area of the ray bundle, although no one has made this connection explicit. In particular, it is possible to calculate the dispersion in the relative fluctuation angle:

$$\sigma(\theta) = 2^{-1/2}\left\langle\left[\delta\vec{\theta}^A - \delta\vec{\theta}^B\right]^2\right\rangle_\theta^{1/2} = \left[C_{\rm gl}(0) - C_{\rm gl}(\theta)\right]^{1/2}$$

$$C_{gl}(\theta) = 16\pi^2 \int_0^\infty k^3 dk \int_0^{\chi_{rec}} P_\phi(k, \tau = \tau_0 - \chi) W^2(\chi, \chi_{rec}) J_0(k\theta \sin_K \chi) d\chi, \quad (5.18)$$

where $\langle\rangle_\theta$ denotes the ensemble average performed over all pairs of photons with a fixed observed angular separation θ and $J_0(x)$ is the Bessel function of order 0.

This is a very useful quantity since it describes in some sense the likelihood of forming caustics. The larger the variance, the more likely are caustics. It should be said that the application of such statistical methods to the study of strong lensing has not occurred yet, however.

Conformally flat spacetimes

In the case of conformally flat spacetimes, the optical scalar equations become:

$$\dot{\theta} + \theta^2 + |\sigma|^2 = \frac{1}{2} R_{\mu\nu} k^\mu k^\nu = \frac{1}{2} T_{\mu\nu} k^\mu k^\nu \quad (5.19)$$

$$\dot{\sigma} + 2\theta\sigma = 0 \quad (5.20)$$

The shear is exactly zero as mentioned before and hence the Dyer-Roeder approximation that the ray bundle's evolution is completely determined by the Ricci focusing term is *exact*. We can write down an equation [142] describing the evolution of the square root of the bundle area, A in this case:

$$(\sqrt{A})'' = -\frac{1}{2H_0^2}(\rho + p)(1+z)^2\sqrt{A} \quad (5.21)$$

where $' \equiv \partial_\nu$ and ν is an affine parameter. This is identical to the FLRW equation. How is this possible, since the conformally flat metrics include models which are not FLRW ? The key lies in the fact that the affine parameter is metric dependent. If we wish to convert to an observable radial coordinate such as redshift, z, then the shear and acceleration of the matter congruence will appear in a non-trivial way to break the symmetry between the different (locally) conformally flat universes, yielding very different observations.

Silent Universes coupled to the optical scalar equations

The discussion on conformally flat models was significantly simplified by the fact that the Weyl tensor was zero. Earlier we discussed the Swiss-Cheese models where it is possible to give a statistical discussion of the Weyl term. Is there any other approach that can be used to estimate the Weyl term ? Of course the heading of this subsection indicates that there is: the use of the Weyl tensor formulated in terms of its electric and magnetic part.

In the silent approximation we go further and drop any pressure terms, and the magnetic part H_{ab} of the Weyl tensor. We also assume that the velocity flow of the matter is irrotational for simplicity. In this case we can write the Weyl term in the optical scalar equations in terms of the electric part of the Weyl tensor. In turn this is coupled to the shear, expansion and energy density of the matter. Thus we are left with 8 equations plus two equations in general to convert the matter and radiation tetrad derivatives to redshift derivatives.

5.5 Conclusions

This concludes our introductory chapter to gravitational lensing, and in particular to the study of area distances in cosmologically non-trivial models. In fact, several new ideas have been included for comparison with existing approaches and include:

• Investigating observations in conformally flat models using the optical scalar equations.

• Using the Silent universe approach to couple the matter exactly to the optical scalar equations. The main problem lies in matching the evolution along matter worldlines and null geodesics. However, this presents perhaps the most sophisticated exact way to study observations in a universe with non-linear inhomogeneities.

Chapter 6

Lensing in Radially Inhomogeneous Universes

6.1 Introduction

Despite the significant advances that have been achieved in studies of lensing, as partially detailed in chapter (5), little progress has been made towards a quantitative understanding of the effect of inhomogeneity on the null cone before and after averaging [1].

To clarify this issue, we ask the question, "take an inhomogeneous universe, average *the geometry* to obtain a FLRW universe and *then* determine the past null cone. Call it C^-_{FLRW}. Now take the original inhomogeneous universe and solving the geodesic equations determine the true past null cone. Call it $C^-{}_{\text{TRUE}}$. Now perform suitable angular averaging on C^-_{TRUE}, to get $\langle C^-_{\text{TRUE}}\rangle$ and compare with C^-_{FLRW}. In particular, is there any way to make the two null cones coincide as a function of redshift ?"

Perhaps from a careful statement of the question it is clear that the null cones are unlikely to be the same, especially as the redshifts in the various models will be different functions of coordinate radius, r, in general. However this has never previously been acknowledged, due principally to the false invocation of the conservation law of photons within the geometric optics limit. The view that the *sky averaged* area of the past null cone in an inhomogeneous universe must coincide with that in the corresponding *matter averaged* FLRW model has been put forward, usually implicitly but also explicitly, many times over the last three decades. Perhaps the origin of the misconception goes back to Weinberg (1976) [216], as will be discussed in depth in the next chapter.

At a basic level, the area of the null cone is determined by the metric properties of spacetime (the null Raychauduri equation) and the application of an averaging procedure is unlikely to lead to a simple conservation law such as photon number, which itself follows from the geometric optics limit of the Einstein-Maxwell equations. This is made all the more dubious by the recognition that there does not exist at present a covariant averaging procedure. Thus by changing coordinates and averaging the geometry in different ways

[1]This work is based on the paper which was done in collaboration with N. Mustapha, C. W. Hellaby and G.F.R. Ellis [213].

one could obtain equally plausible but different, FLRW spacetimes with different null cone properties. For example, in the usual $3+1$ comoving coordinate splitting, the "natural" averaging is over the spatial sections defined orthogonal to the four velocity. This is the averaging used in astrophysics and employed in this chapter. However, using null-cone coordinates (see e.g. [207]) where the "time"-coordinate identifies individual null-cones, the "natural" averaging is over the past-null cones, giving completely different results.

However, the depth of the belief that the averaged area distance in an inhomogeneous universe must be equal to the "corresponding" [2] FLRW area distance should not be underestimated. The usual counter-argument goes roughly as follows:

When a bundle of rays goes through an underdensity (relative to the "corresponding" background FLRW model), the area of the bundle increases over the fiducial FLRW bundle. However, because the matter averages out, there must be associated matter overdensities which cause the bundle of rays to be focussed by the tidal shearing, decreasing the area. When averaging is performed the two effects cancel out leaving, on average, the area distance of the background FLRW model. While the basic processes described are correct, the belief that matter averaging translates to wavefront averaging is incorrect.

The existence of caustics makes it easier to see the flaw in the above argument. While it is true that prior to the conjugate point (formation point of the caustic), the area of the ray bundle is decreasing due to the convergence and shearing effects, once the bundle moves through the conjugate point, the area begins *increasing* again due to the divergence of the geodesic spray. Meanwhile in the underdense regions, there is always divergence. This leads to the conjecture that at high redshift, $z > 1$, when caustics are common, not only will the area distance differ from the FLRW area distance, but it will be much larger due to the extra area of the multi-sheeted caustics. This is discussed in the following chapter in detail.

The aim of this chapter, in contrast, is to explicitly provide a model proving the non-equality of the averaged area-distance and the FLRW area-distance, and hence the failure of the photon-conservation argument. Since it is a counter-proof, the ideal model should be simple and should be one in which averaging can be performed easily. This is done using the Lemaître-Tolman-Bondi (LTB) model. Because the model is spherically-symmetric, the angular averaging is already performed, while the single degree of freedom in the radial direction makes matter averaging relatively simple. Despite this there are significant subtleties which are detailed in the following pages and the associated paper [213].

We construct an exact inhomogeneous model and its FLRW approximation, and compare the area distance-redshift and density-redshift relations in the two. To do so we examine Lemaître-Tolman-Bondi (LTB) spherically symmetric dust solutions [221, 222, 223], where exact integrations of the field equations are available for the past light cone of observers at the central position. Although the lensing that occurs for this central observer is purely radial, we find the inhomogeneity has a tangible effect on observational relations.[3]

This chapter does not generically establish the magnitude of the effect, precisely because the high-symmetry geometry considered here precludes formation of caustics and the

[2] Again there is no unique way to assign the corresponding FLRW model !

[3] Radial lensing is a spherically symmetric distortion of the null cone compared with an FLRW model, resulting in a uniform delay of the wavefront. There is no image distortion, no dependence of magnification or time delay on direction, and no multiple imaging, but our results show its effects are observable.

consequent fractal-like structure of the real light cone.

In developing the results of this chapter and [213], we solve one of the problems that has made analysis of observations in Lemaître-Tolman-Bondi solutions difficult, namely the problem of precisely locating the past light cone of the chosen central event P, by use of a special choice of radial coordinate that ensures a very simple form for the past light cone of P in these inhomogeneous space-times. This technical development has other uses in terms of analysing observational relations in these models.[4]

6.2 Method and Program

We select the simplest inhomogeneous solution of the Einstein Field Equations; the Lemaître-Tolman-Bondi (LTB) model which is spherically symmetric, but radially inhomogeneous, with a dust equation of state.

The question we are raising is whether the area of an averaged wavefront we receive at our observatory in an inhomogeneous universe is the same as the area of a wavefront in a smoothed version of that universe. To clarify this issue our strategy is to:

- Select the most natural generalisation of the Einstein-de Sitter models commonly used in studies of observations and describe data on the null cone (i.e. a parabolic LTB model).
- Find the FLRW limit of this inhomogeneous universe in an appropriate coordinate system.
- Average the lumpy universe in a natural way and fit it correctly to a FLRW model.
- Compare area distances in the lumpy universe and its smoothed average.

6.2.1 The Inhomogeneous Model

The Integrated Field Equations

We choose the parabolic LTB model which is the natural generalisation of the $\Omega = 1$ dust FLRW model. This model is characterised by the mass within comoving radius r, $M(r)$, and so- called 'bang-time' function $t_B(r)$ describing the locus of the initial spatial hypersurface (that is, the local time of the big bang), up to a coordinate freedom.

In normalised comoving coordinates the metric after solving the off- diagonal EFE is

$$ds^2 = -dt^2 + (R')^2 dr^2 + R^2(d\theta^2 + \sin^2\theta d\Phi^2) \tag{6.1}$$

where $R'(t,r) = \partial R(t,r)/\partial r$.

The time curves are irrotational and, for comoving dust ($p = 0$), are necessarily geodesics because of momentum conservation. The spatial sections are flat because if we choose $r = R(t_0, r)$ then $R'(t_0, r) = 1$ and we find that the 3-spaces have metric $d\sigma^2 = dr^2 + r^2 d\Omega^2$ and hence are flat.

The areal radius, $R = R(t,r)$ in the Lemaître-Tolman-Bondi metric, *is* the area of the intersection of our past null cone with past spacelike time surfaces (in this case spheres)

[4] For a slightly different analysis of LTB spacetimes, based on null cone coordinates, see [224], and for a consideration of observations away from the centre of symmetry see [225].

once specification has been made that $R \propto r$ for small r. In the parabolic case R is given explicitly by the solution to the equation of motion

$$\dot{R}(t,r) = \sqrt{\frac{M(r)}{R(t,r)}} \tag{6.2}$$

obtained from the 11, 22 and 33 components of the EFE, where $\dot{}$ denotes the derivative with respect to t; i.e.

$$R(t,r) = \left[\frac{9M(r)}{2}(t - t_B(r))^2\right]^{1/3} \tag{6.3}$$

where t is cosmic time whilst M and t_B are both functions of coordinate radius r only. It follows immediately that

$$R'(t,r) = \frac{3}{2R^2(t,r)}\left[M'(r)(t - t_B(r))^2 - 2M(r)(t - t_B(r))t'_B(r)\right] \tag{6.4}$$

where $'$ denotes the derivative with respect to radial coordinate r.

The 00 field equation gives

$$4\pi\rho(t,r) = \frac{M'(r)}{R^2(t,r)R'(t,r)}\,. \tag{6.5}$$

The Solution on the Null Cone

Since we are interested in observations on the null cone we must project onto it by specifying the unique relationship between r and t. On radial null geodesics, $ds^2 = 0 = d\theta^2 = d\Phi^2$, so from (6.1), if the past light cone of the event $(t = t_0, r = 0)$ is given by $t = \hat{t}(r)$, then that light cone is described by

$$dt = \pm R'(\hat{t}(r), r)dr\,. \tag{6.6}$$

The coordinate freedom in the LTB metric is a rescaling of the radial coordinate $r \to \tilde{r} = \tilde{r}(r)$. If we choose r so that

$$R'(\hat{t}(r), r) = 1\,, \tag{6.7}$$

then on the past light cone $dt = -dr$, so that the incoming light rays at the event $(t = t_0, r = 0)$ are given by

$$\hat{t}(r) = t_0 - r\,. \tag{6.8}$$

So this gauge choice, in contrast to other work done on observations in the LTB model, locates the null cone at one instant of time in its simplest possible form and makes our programme analytically solvable. On this light cone, putting (6.8) in (6.4) and using (6.7),

$$R'(\hat{t}(r), r) = \frac{3}{2R^2(\hat{t}(r), r)}[M'(r)\tau(r)^2 + 2M(r)\tau(r)\tau'(r) + 2M(r)\tau(r)] = 1 \tag{6.9}$$

where we have defined

$$\tau(r) = t_0 - r - t_B(r)\,. \tag{6.10}$$

The function $\tau(r)$ can be interpreted as proper time from the bang surface to our past null cone along the particle worldlines. We can set t_0 to be the time since the bang at the observer $(r = 0)$ by choosing $t_B(0) = 0$ (so $\tau(0) = t_0$).

It is important to realise that evaluating $R'(t,r)$ on the null cone $t = \hat{t}(r)$ is not the same as differentiating $\hat{R}(r) = R(\hat{t}(r), r)$ with respect to r. In fact, by evaluating (6.3) on the null cone, $\hat{R}$ is given by

$$\hat{R} = R(\hat{t}(r), r) = \left[\frac{9M(r)}{2}\tau(r)^2\right]^{1/3} \tag{6.11}$$

which means that its derivative is given by

$$\frac{d\hat{R}}{dr} = \frac{d}{dr}[R(\hat{t}(r), r)] = \frac{3}{2R^2(\hat{t}(r), r)}[M'(r)\tau(r)^2 + 2M(r)\tau(r)\tau'(r)]\,. \tag{6.12}$$

Combining the above equation with the constraint (6.9) gives a first order differential equation for $\hat{R}$.

$$\frac{d\hat{R}(r)}{dr} + \frac{3M(r)\tau(r)}{\hat{R}(r)^2} = 1\,. \tag{6.13}$$

In summary, with our choice of coordinates we have recast the flat LTB model in a form that allows us to locate the past null cone with ease. This has left us with one freedom to choose an arbitrary function of r. We could choose τ (or M) and substitute (6.11) into (6.13). Solution of this differential equation would determine $\hat{R}$ and thus any other quantity. If we instead decide to choose $\hat{R}$, that is, the area of the wavefront, then the model is trivially and fully specified by (6.11) and (6.13). It follows that

$$\tau(r) = \frac{2\hat{R}(r)}{3} \Big/ \left(1 - \frac{d\hat{R}(r)}{dr}\right) \tag{6.14}$$

and

$$M(r) = \frac{\hat{R}(r)}{2}\left(1 - \frac{d\hat{R}(r)}{dr}\right)^2 \qquad (M(0) = 0)\,. \tag{6.15}$$

To obtain the results of the next section, we will choose $\hat{R}$ and find its derivative. This will then determine $\tau(r)$ (equivalently $t_B(r)$) and $M(r)$ by the above two equations. The flat LTB model will thus be fully specified in these coordinates and one could then propagate the data off the null cone by the comoving assumption.
The density on the null cone $\hat{\rho}(r)$ is found by evaluating (6.5) on the null cone:

$$4\pi\hat{\rho}(r) = \frac{M'(r)}{\hat{R}(r)^2} \tag{6.16}$$

and its value at the origin depends on the time as characterised by the Hubble constant;[5]

$$\Omega = \frac{4\pi\rho}{H^2} = 1 \quad \Rightarrow \quad \rho_0 = \frac{H_0^2}{4\pi}, \quad t_0 = \frac{2}{3H_0}\,. \tag{6.17}$$

[5]Since measurements of the Hubble constant are taken at about $z < 1$, we can take this to determine the age of the universe, t_0, at the central observer.

Redshifts

It is of some importance that we state the relevant quantities in terms of redshifts. To do this, we use the fact that in the geometric optics limit, for two light rays emitted on the worldline at r_{em} with time interval $\delta t_{em} = t^+(r_{em}) - t^-(r_{em})$ and observed on the central worldline with time interval $\delta t_{ob} = t^+(0) - t^-(0)$

$$1 + z = \frac{\delta t_{ob}}{\delta t_{em}}. \tag{6.18}$$

The past radial null geodesics are given by

$$dt = -R'(t,r)dr\,,$$

so for an observer on a nearby worldline, the time interval changes by

$$d(\delta t) \;=\; dt^+ - dt^- \;=\; [R'(t^-,r) - R'(t^+,r)]\,dr \;=\; -\frac{\partial}{\partial t}\,[R'(t,r)]\,\delta t dr\,.$$

Thus

$$d\ln\delta t = -\frac{\partial}{\partial t}\,[R'(t,r)]\,dr$$

which means that the redshift, given by (6.18), is

$$\ln(1+z) = \int_0^{r_{em}} \dot{R}'(\hat{t},r)dr \tag{6.19}$$

where $\hat{t}(r)$ is the equation of the null cone.[6] To calculate $\dot{R}'(t,r)$ we differentiate (6.3) with respect to t

$$\dot{R}' = \frac{\dot{R}}{3}\left[\frac{t_B'}{t-t_B} + \frac{M'}{M}\right], \qquad \dot{R} = \left[\frac{4M}{3(t-t_B)}\right]^{1/3}. \tag{6.20}$$

Since $\hat{t}(r) = t_0 - r$ when we choose $R' = 1$ on the null cone, $\dot{R}'(\hat{t},r)$ is given by

$$\dot{R}'(\hat{t},r) = \frac{1}{3}\left[\frac{4M}{3\tau}\right]^{1/3}\left[\frac{M'}{M} - \frac{1+\tau'}{\tau}\right]. \tag{6.21}$$

After some manipulation of the above expression substituted into (6.19), we find that

$$\ln(1+z) = \left(\frac{4M}{3\tau}\right)^{1/3} - \frac{1}{3}\int_0^{r_{em}}\left(\frac{4M}{3\tau^4}\right)^{1/3} dr\,. \tag{6.22}$$

Using (6.14) and (6.15) this equation may be written as

$$\ln(1+z) = \left(1 - \frac{d\hat{R}}{dr}\right) - \frac{1}{2}\int_0^{r_{em}}\left(1 - \frac{d\hat{R}}{dr}\right)^2 / \,\hat{R}\,dr \tag{6.23}$$

so we can now determine the redshift-area distance relation.

[6] The standard formula $1 + z = (u^\mu k_\mu)_{em}/(u^\mu k_\mu)_{ob}$ is not useful in this gauge since $k^\mu = (R', -1, 0, 0) = (1, -1, 0, 0)$ is not affine, though it is tangent to the past null cone.

6.2.2 The Friedmann-Lemaître limit

The characterisation of the FLRW limit is that the bang time surface is simultaneous. So $t_B(r) = t_{B\text{FLRW}} = \text{constant}$; from whence

$$R_{\text{FLRW}}(t,r) = \left[\frac{9}{2}M_{\text{FLRW}}(r)(t - t_{B\text{FLRW}})^2\right]^{1/3}, \quad R'_{\text{FLRW}}(t,r) = \frac{3}{2R^2_{\text{FLRW}}}M'_{\text{FLRW}}(r).(t - t_{B\text{FLRW}})^2. \tag{6.24}$$

The freedom left here in $M_{\text{FLRW}}(r)$ is just essentially the coordinate freedom, corresponding to the freedom of choice of r. The above relations determine the FLRW density

$$\rho_{\text{FLRW}}(t) = \frac{1}{6\pi(t - t_{B\text{FLRW}})^2} \tag{6.25}$$

which is spatially homogeneous as required, unaffected by $M_{\text{FLRW}}(r)$. It is usual to set $t_{B\text{FLRW}} = 0$. We do not have a freedom to rescale the density by a constant because this is the critical density case.

As we would eventually like to compare our LTB model as chosen above to an underlying FLRW model, it is appropriate to write the FLRW limit in the same kind of coordinate system. Consider light rays coming in to the event $(t = t_1, r = 0)$ in a FLRW model. When we choose coordinates for which $R'_{\text{FLRW}}(t,r) = 1$ on the null cone, the past null cone can be located by $\hat{t} = t_1 - r - t_{B\text{FLRW}} = t_1 - r$. (We use t_1 rather than t_0 here, as we will need to distinguish LTB and FLRW values later on. As a limit of the flat LTB model in these coordinates, the FLRW form of $M(r)$ is obtained from setting $\tau = t_1 - r$ in (6.9). This yields

$$M_{\text{FLRW}} = 6\left[t_1^{1/3} - (t_1 - r)^{1/3}\right]^3 \qquad (M_{\text{FLRW}}(0) = 0) \tag{6.26}$$

$$\hat{R}_{\text{FLRW}} = 3\left[t_1^{1/3} - (t_1 - r)^{1/3}\right](t_1 - r)^{2/3} \qquad (R_{\text{FLRW}}(0) = 0). \tag{6.27}$$

We note that this in conjuction with (6.24) implies that

$$R_{\text{FLRW}}(t,r) = 3\left[t_1^{1/3} - (t_1 - r)^{1/3}\right]t^{2/3} \quad , \quad R'_{\text{FLRW}} = \frac{3M'_{\text{FLRW}}}{2R^2_{\text{FLRW}}}t^2 = \frac{t^{2/3}}{(t_1 - r)^{2/3}}. \tag{6.28}$$

The RW metric that results is, from (6.28) and (6.1),

$$ds^2 = -dt^2 + t^{4/3}\left\{\frac{1}{(t_1 - r)^{4/3}}dr^2 + 9\left[t_1^{1/3} - (t_1 - r)^{1/3}\right]^2 d\Omega^2\right\}. \tag{6.29}$$

These coordinates are singular at the particle horizon, $r = t_1$ (when the past null cone of $t = t_1$ runs into the initial singularity). Thus they are valid for $0 \le r < t_1$. The FLRW redshift-distance formula can be obtained by inserting the FLRW forms of $M(r)$ and $\tau(r)$ into equation (6.22). That is

$$z(r) = \left(\frac{t_1}{t_1 - r}\right)^{2/3} - 1 \qquad \Longleftrightarrow \qquad r(z) = t_1\left[\frac{(1+z)^{3/2} - 1}{(1+z)^{3/2}}\right]. \tag{6.30}$$

6.2.3 Averaging and Fitting

We want to compare and contrast total areas of wavefronts at given redshifts of an inhomogeneous model to that of the corresponding FLRW model of density equal to the inhomogeneous density perfectly smoothed. This must be done with respect to the inhomogeneous metric because physically the smoothing does not occur.

Perhaps the crucial part of our analysis is that we ensure that we compare with the FLRW model with the correct average density. We define the *average* or *background* FLRW model to be the one that matches on at the particle horizon where $\tau = 0$, $r = r_\Sigma$, using the Darmois-Israel boundary conditions [226, 227]. Matching first and second fundamental forms of this timelike (comoving) boundary surface Σ gives

$$R_{\rm LTB}|_\Sigma = R_{\rm FLRW}|_\Sigma \tag{6.31}$$

and the background model must be parabolic if the inhomogeneous one is; or vice versa.

The matching must hold over all of Σ; that is, at all times – so

$$\dot{R}_{\rm LTB}\Big|_\Sigma = \dot{R}_{\rm FLRW}\Big|_\Sigma \tag{6.32}$$

and thus, by (6.2),

$$M_{\rm LTB}|_\Sigma = M_{\rm FLRW}|_\Sigma \, . \tag{6.33}$$

Thus it is sufficient to match the masses at Σ, and synchronise the starting times (bang times) when $R_{\rm LTB}|_\Sigma = 0 = R_{\rm FLRW}|_\Sigma$. In general, we do not expect the FLRW radial coordinate on Σ ($r_{\rm FLRW}|_\Sigma$) to be the same as the LTB one there ($r_{\rm LTB}|_\Sigma = r_\Sigma$) since the coordinate condition $\hat{R}' = 1$ holds on the null cone, whose locus is model dependent.

For a parabolic LTB model with metric (6.1) and density given by (6.5), the background density $\rho_{\rm FLRW}$ is the same as that obtained by integrating over constant time slices.

$$\begin{aligned}
\langle\rho\rangle_{t_0,r_\Sigma} &= \left(\int_0^{2\pi}\int_0^{\pi}\int_0^{r_\Sigma} \rho\sqrt{^3g}\; dr d\theta d\Phi\right) \Big/ \left(\int_0^{2\pi}\int_0^{\pi}\int_0^{r_\Sigma} \sqrt{^3g}\; dr d\theta d\Phi\right) \\
&= \left(\int_0^{r_\Sigma} \frac{M'}{R^2R'}\sqrt{R'^2R^4}dr\right) \Big/ \left(4\pi\int_0^{r_\Sigma}\sqrt{R'^2R^4}dr\right) \\
&= \frac{3}{4\pi}\frac{M(r_\Sigma)}{[R(r_\Sigma,t_0)]^3} \\
&= \frac{1}{6\pi(t_0 - t_B(r_\Sigma))^2} = \rho_{\rm FLRW}
\end{aligned} \tag{6.34}$$

where equation (6.3) was used.

One important point that must be made here is that a covariant averaging procedure does not exist as yet. We have used here an averaging method which is 'natural' for the comoving synchronous coordinates which lead to a $3+1$ foliation of spacetime. However, the same model in different (for example observational) coordinates would suggest a different averaging procedure which could conceivably yield different results. Therefore the claim [216] that the wavefront areas obtained in the inhomogeneous model and the averaged model are the same already seems highly unlikely.

6.3 Results

We use geometric units such that $G = c = 1$. If we choose a unit of time T_G *seconds* to be 1 geometric time unit (*gtu*), then the geometric units of length, mass, density, etc. are fixed by 1 $glu = L_G = cT_G$ *metres*, 1 $gmu = M_G = (c^3/G)T_G$ *kg*, 1 $gmu\ glu^{-3} = \rho_G = (1/G)T_G^{-2}$ $kg\ m^{-3}$. For the purposes of this chapter, we want units suitable to cosmological scales, so we specify that one cosmological time unit, 1 *ctu*, is ten billion years – of the order of the age of the universe. This gives us

Cosmological Geometric Units

	Time	Length	Mass	Density
Cosmological	1 *ctu*	1 *clu*	1 *cmu*	1 $cmu\ clu^{-3}$
SI	$3.156 \times 10^{17}\ s$	$9.461 \times 10^{25}\ m$	$1.275 \times 10^{53}\ kg$	$1.505 \times 10^{-25}\ kg/m^3$
Astronomical	10 *Gyr*	3.066 *Gpc*	$6.409 \times 10^{22}\ M_\odot$	$1.505 \times 10^{-28}\ g/cc$

The first subsection (6.3.1) gives a very simple model which satisfies the criteria for a reasonable cosmological model (with the classical Copernican principle dropped) and which provides a *proof that there exist physically reasonable density behaviours which lead to a nonzero magnification or shrinking.* It is obvious that averaging over the sky will not remove this effect since the model is already spherically symmetric. The second model (6.3.1) does the same, but is smoother at the origin and displays interesting behaviour in redshift space.

These two models are obtained by choosing the observer area distance function, which is the easiest way of solving this problem.

6.3.1 Form of Perturbation and General Results

It is not easy to choose a form of area distance function for the inhomogeneous model which results in reasonable physical behaviour. So instead we choose it in the form of a 'perturbation' from a flat Friedmann model; that is,

$$\hat{R}(r) = \hat{R}^u_{\text{FLRW}}(r)(1 + \delta(r)) \tag{6.35}$$

where, from (6.27), $\hat{R}^u_{\text{FLRW}}(r) = 3[t_u{}^{1/3} - (t_u - r)^{1/3}](t_u - r)^{2/3}$ is the area function of an *underlying* FLRW model of age $t_1 = t_u$. (This 'underlying' FLRW model is a mathematical device with no physical significance. It can not be considered a background or average model since we have not restricted $\delta(r)$ to average out to zero in any sense.) In principle, one should choose a density function and then determine the area distance function from it or risk the possibility of assuming the result. However, if we can show that the above choice of $\hat{R}$ leads to a density profile with reasonable physical behaviour, this would suffice – since if we had initially chosen that density function, it would lead to an $\hat{R}$ as chosen above. We will show that this is indeed the case and also indicate that the model is free of shell crossings.[7]

[7]The necessary and sufficient conditions for there to be no shell crossings anywhere or at any time in the evolution of a flat model with $R' > 0$ are that $M(r)$ be an increasing and $t_B(r)$ a decreasing function. They were found (for all LTB spacetimes) by Hellaby and Lake [228].

Obviously $\hat{R}(r)$ is zero at the same places as $\hat{R}^u_{\text{FLRW}}(r)$, i.e. at $r = 0$ and at $r = t_u$. For this form of perturbation, in terms of the convenient parametrisation $v = r/t_u$, we find

$$\begin{aligned}
X &\equiv \left[\frac{1}{(1-v)^{1/3}} - 1\right] && (6.36)\\
M &= \frac{3}{2}t_u X(1-v)(1+\delta)[2X(1+\delta) - \delta - 3t_u X(1-v)\delta']^2 && (6.37)\\
\tau &= \frac{2t_u X(1-v)(1+\delta)}{[2X(1+\delta) - \delta - 3t_u X(1-v)\delta']} && (6.38)\\
t_B &= t_0 - r - \tau && (6.39)\\
8\pi\hat{\rho} &= \frac{2X(1+\delta) - \delta - 3t_u X(1-v)\delta'}{9[t_u X(1-v)(1+\delta)]^2}\{2X(3+4\delta)(1+\delta) - \delta(1+\delta) - 3t_u X(1-v)(5+6\delta) \\
&+ 36t_u X^2(1-v)(1+\delta)\delta' - 9{t_u}^2 X^2(1-v)^2[2(1+\delta)\delta'' + \delta'^2]\} && (6.40)\\
\frac{d\ln(1+z)}{dz} &= \{4X(1+2\delta)(1+\delta) - \delta^2 - 6t_u X(1-v)(2+3\delta)\delta' \\
&+ 36t_u X^2(1-v)(1+\delta)\delta' - 9{t_u}^2 X^2(1-v)^2[2(1+\delta)\delta'' + \delta'^2]\} \,/\, [6t_u X(1-v)(1+\delta)] && (6.41)
\end{aligned}$$

If $\delta(0) \neq 0$ we find the unphysical limits $\tau(0) = 0$ and $\hat{\rho}(0) = \infty$. Thus we set $\delta(0) = 0$, obtaining the following limiting values

$$\begin{aligned}
M(0) &= \left.\frac{2(1 - 3t_u\delta'(r))^2 r^3}{9{t_u}^2}\right|_{r\to 0} = 0 && (6.42)\\
\tau(0) &= \frac{t_u}{1 - 3t_u\delta'(0)} && (6.43)\\
8\pi\hat{\rho}(0) &= \frac{4(1 - 3t_u\delta'(0))^2}{3{t_u}^2} && (6.44)\\
\left.\frac{d\ln(1+z)}{dr}\right|_{r=0} &= \frac{2(1 - 3t_u\delta'(0))}{3t_u} && (6.45)
\end{aligned}$$

and

$$\begin{aligned}
M(t_u) &= 6t_u(1 + \delta(t_u))^3 && (6.46)\\
\tau(t_u) &= (t_u - r)|_{r\to t_u} = 0 && (6.47)\\
8\pi\hat{\rho}(t_u) &= \left.\frac{4(3 + 4\delta(r))}{9(t_u - r)^2}\right|_{r\to t_u} = \infty && (6.48)\\
\left.\frac{d\ln(1+z)}{dr}\right|_{t_u} &= \left.\frac{2(1 + 2\delta(r))}{3(t_u - r)}\right|_{r\to t_u} = \infty\,. && (6.49)
\end{aligned}$$

From numerical experimentation we concluded, in order to avoid shell crossings, that $\delta(r)$ must remain sufficiently far away from zero over most if not the entire range of r, and certainly near $r = t_u$. We want the proper time from the bang surface to the null surface on the central worldline to be the 'true' age of the universe; that is, we want it to be t_0, the time at the origin of the LTB model. By setting $\tau(0) = t_0$ in (6.43), the age of the underlying model is determined

$$t_u = t_0\,(1 - 3\,t_1\,\delta'(0))\,. \qquad (6.50)$$

The parameter t_u is the r-coordinate value at which the null cone of the LTB model intersects the bang. We will average quantities on this scale; that is to say, we shall take $r_\Sigma = t_u$.

We match this inhomogeneous universe to a flat FLRW model at the surface r_Σ by equating the masses and bang times at that point. This then determines the time $t_1 = t_b$ in the *background* FLRW model which we will use for our comparison. At $r = t_u$, $\hat{R} = 0$ and (6.46) shows that at this point,

$$M(t_u) = M^u_{\mathrm{FLRW}}(t_u)\,(1+\delta(t_u))^3 = 6t_u\,(1+\delta(t_u))^3 .$$

In the background FLRW model the value of the mass at Σ is $6t_b$ and this is what we have to match the inhomogeneous mass to. This gives us a value of the age for the background flat FLRW model of

$$t_b = t_u\,(1+\delta(t_u))^3 . \tag{6.51}$$

A Regular Model which Exhibits Shrinking and Magnification

The following simple example is physically well behaved, being free of shell crossings at all times in its evolution for $r \le t_u$. Since $t'_B \ne 0$ at the origin, the model is not as smooth there as one would like, but there are no physical problems. We choose $\delta(r)$ for our first model, LTB1, to be

$$\delta(r) = -\frac{1}{5}\sin\left(\frac{0.8\pi r}{t_u}\right) . \tag{6.52}$$

When we set $\tau(0) = t_0 = 1$ then

$$t_u = t_0\left(1+\frac{12\pi}{25}\right) \tag{6.53}$$

and the age for the background flat FLRW model, after matching the masses, is

$$t_b = t_u\left(1-\frac{\sin 0.8\pi}{5}\right)^3 . \tag{6.54}$$

The calculation of the redshift was done by a numerical quadrature of (6.23).

It is important to plot these quantities in terms of the observable quantity z for two reasons. First of all, in the transformation $r \to z$, the possibility exists that the area distances of the flat and inhomogeneous models might transform into each other. Secondly, under certain circumstances the redshift becomes disordered with distance and unexpected behaviour might occur, as the following model illustrates.

A Regular Model with Multivalued Observations

As an illustration of how different the physical quantities plotted against radial coordinate r as opposed to those same quantities plotted against redshift z may appear, we present here an LTB model for which the redshift becomes disordered with distance at some points and then ordered again at later points.

The universe is chosen as above but with a 'perturbation function' of

$$\delta(r) = -\frac{1}{4}\sin\left(\frac{0.75\pi r}{t_u}\right) - \frac{1}{4^6}\left(\frac{11}{\pi}+\frac{3}{2}\right)\left[1-\cos\left(\frac{4\pi r}{t_u}\right)\right] . \tag{6.55}$$

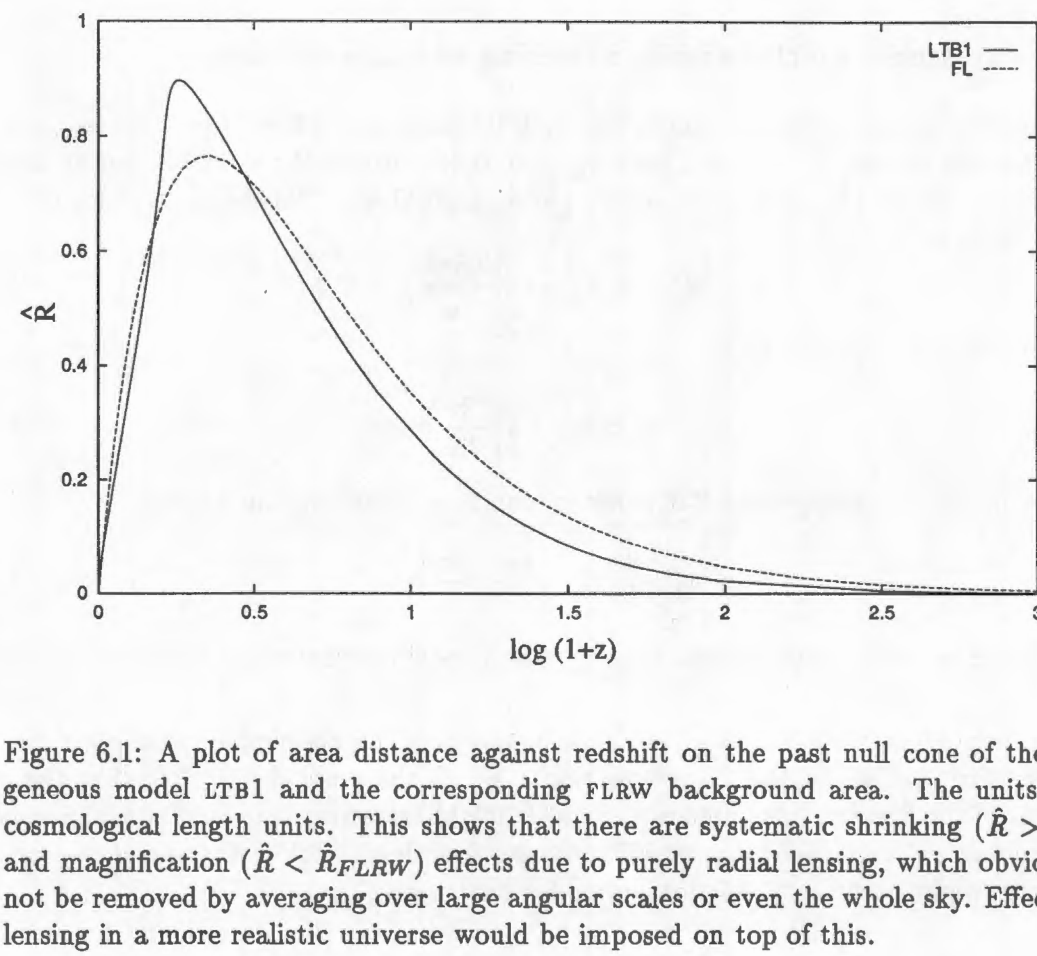

Figure 6.1: A plot of area distance against redshift on the past null cone of the inhomogeneous model LTB1 and the corresponding FLRW background area. The units of $\hat{R}$ are cosmological length units. This shows that there are systematic shrinking ($\hat{R} > \hat{R}_{FLRW}$) and magnification ($\hat{R} < \hat{R}_{FLRW}$) effects due to purely radial lensing, which obviously cannot be removed by averaging over large angular scales or even the whole sky. Effects of true lensing in a more realistic universe would be imposed on top of this.

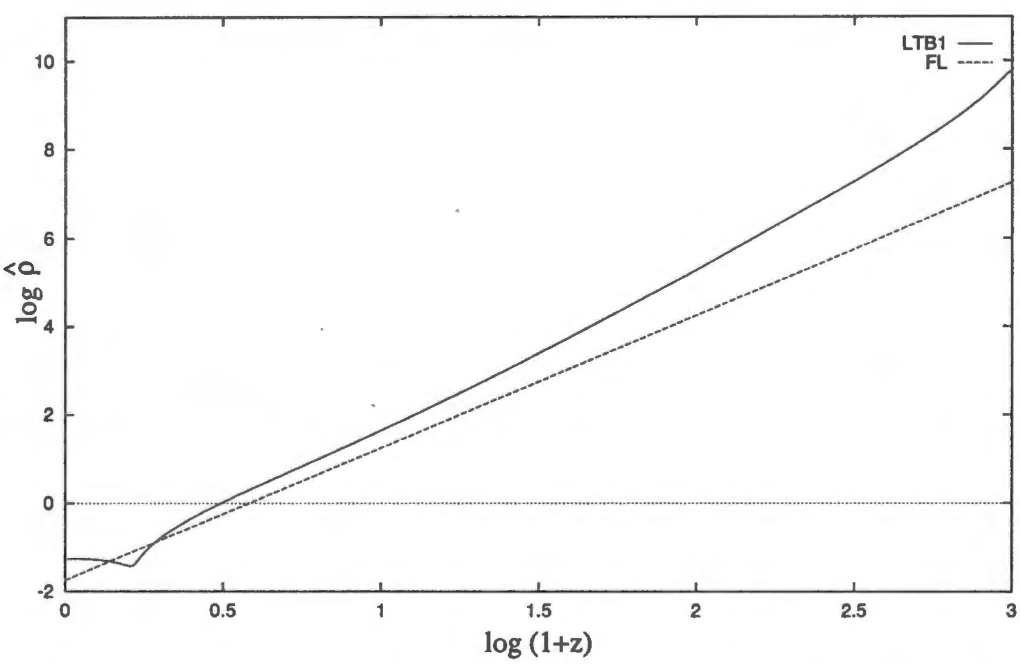

Figure 6.2: The density of matter on the *past null cone* (that is, what would actually be observed) in the models of FIG. 6.1, LTB1 and its corresponding background FLRW model. The units are cosmological density units ($cmu\ clu^{-3}$). When comparing with FIG. 6.1, we see that roughly speaking, magnification occurs for objects in or just beyond an overdense region, and shrinking occurs for objects in or just beyond an underdensity.

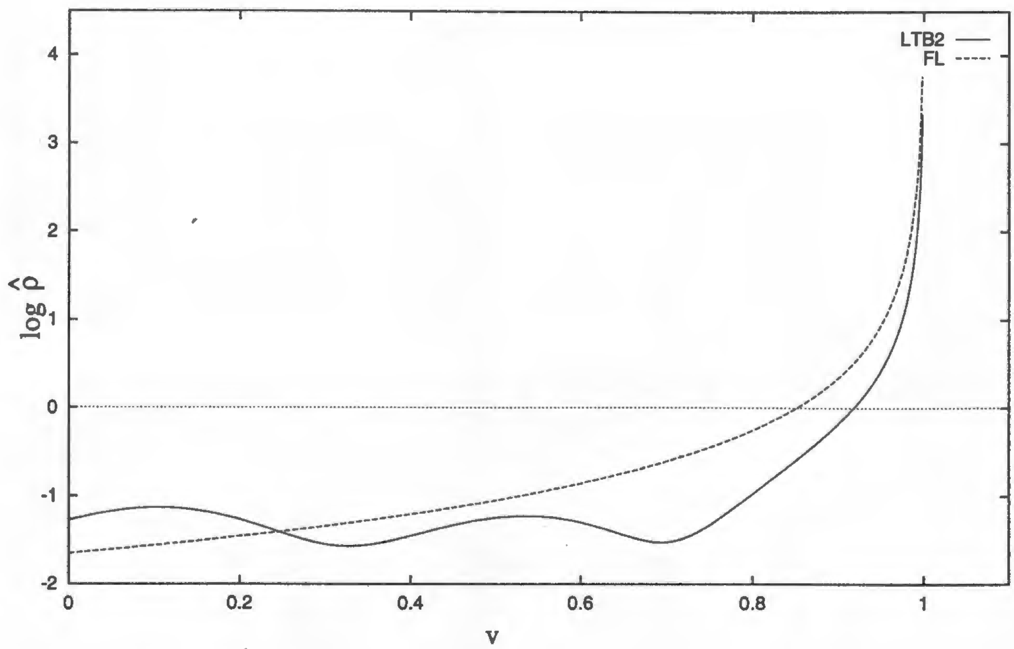

Figure 6.3: The densities for the second LTB model and its background FLRW model ($\hat{\rho}$ and $\hat{\rho}_{FLRW}$, in cosmological units) on the past null cone. Again, the inhomogeneous profile vs r appears quite acceptable.

This model, which we call LTB2, is also free of shell crossings at any time for $r \leq t_u$ and has a completely smooth and regular origin (where $t'_B = 0$). Setting $\tau(0) = t_0 = 1$ once again gives

$$t_u = t_0 \left(1 + \frac{9\pi}{16}\right) \tag{6.56}$$

and

$$t_b = t_u \left(1 - \frac{\sin 0.75\pi}{4}\right)^3 . \tag{6.57}$$

This model provides a good illustration of why one has to be careful in ascribing physical behaviour in a certain coordinate system. Viewed as functions of r, $\hat{R}$ and $\hat{\rho}$ have fairly standard behaviour, but viewed in terms of the observable quantity z, the density and area distance become multivalued. Hence, three objects with the same intrinsic luminosity located at different distances appear at the same z, with three different apparent luminosities (or area distances).

Our numerical experiments indicate that the redshift on the light cone is most sensitive to perturbations in the vicinity of the maximum in $\hat{R}(z)$. All our models in which $dz\,/\,dr$ became negative did so in this region. The looping behaviour in the $\hat{R}$ vs $\log(1+z)$ plot occurs when the maximum and minimum in the $\log(1+z)$ vs r graph bracket the maximum in the $\hat{R}$ vs r graph. Similarly, perturbations more easily generate a maximum and minimum in the $\hat{\rho}(z)$ near the maximum in $\hat{R}$, hence the loop in that graph.

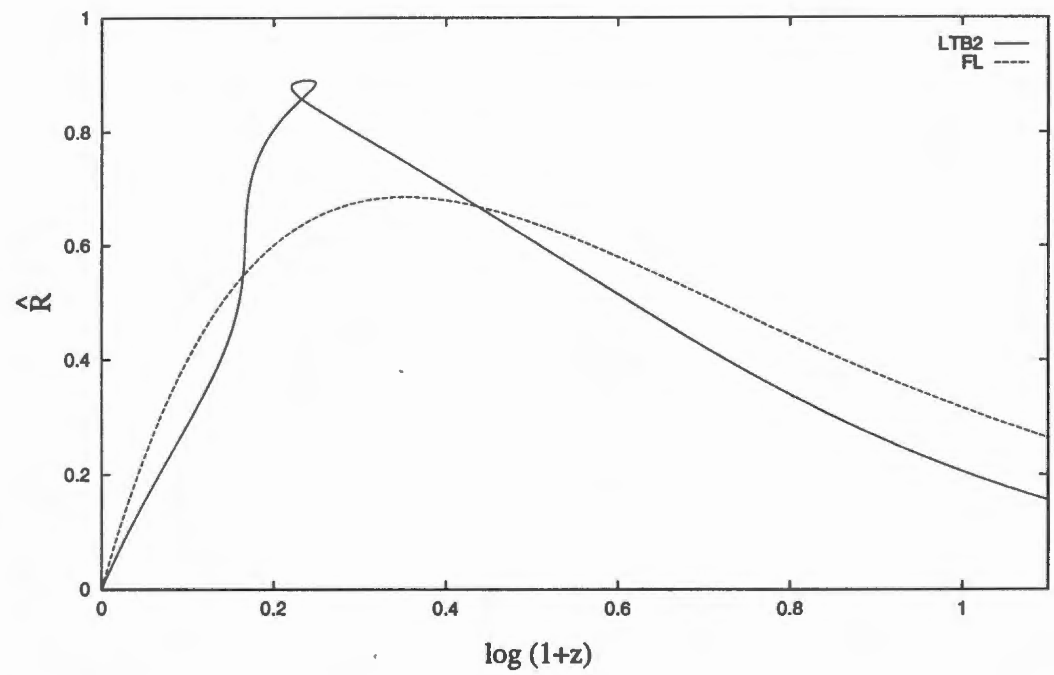

Figure 6.4: A plot of area distance against redshift for model LTB2 and the background FLRW model. The interesting point to note is that at some redshifts the area distance in the LTB model is multivalued.

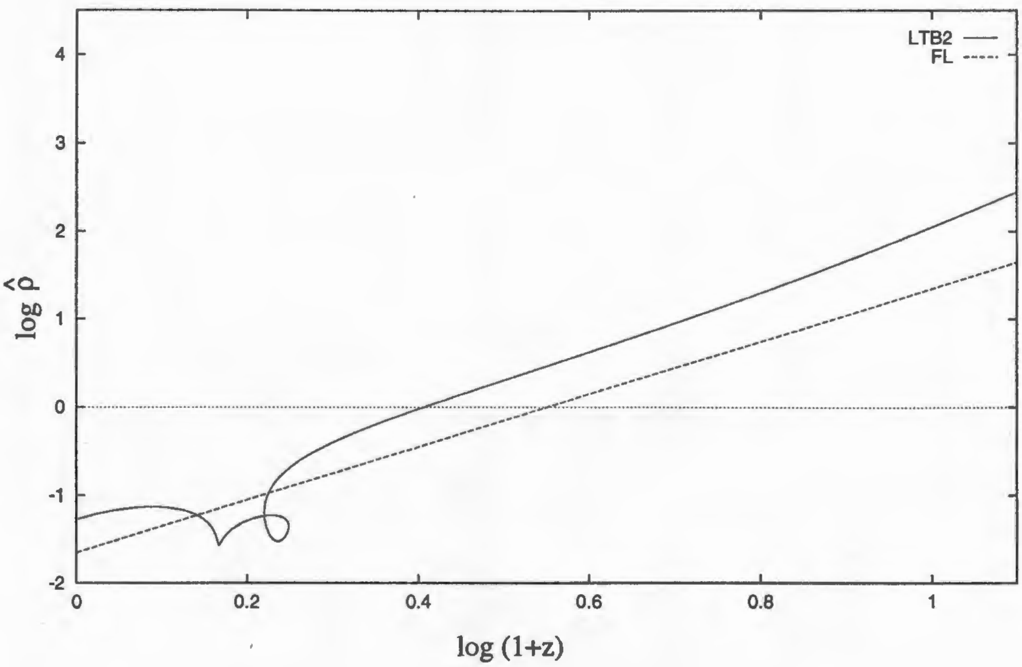

Figure 6.5: The densities ($\hat{\rho}$ and $\hat{\rho}_{FLRW}$) on the past null cones vs z for model LTB2 and its background FLRW model. Note the quaint 'looping' behaviour.

6.4 Conclusions

The general belief that photon conservation implies that the total area of an incoming wavefront must be the same as in the background, matter-averaged, FLRW model has been disproved in this chapter. The spherically symmetric model used here is simple but effective, since averaging over direction cannot change the results. In more realistic models of the lumpy universe this effect will still be present, and we expect full gravitational lensing to occur, resulting in more significant deviations from the FLRW formula.

This investigation used a parabolic LTB model, where the areal radius $\hat{R}$ is also the area distance of the 2-sphere wavefronts of the past null cone. The density in the LTB model is averaged to give a background Einstein-de Sitter ($\Omega = 1$) model, and it is tested against this model. Although there exists no covariant way to perform this averaging, we use the 'natural' one defined by the use of junction conditions, here equivalent to the one used in astrophysical problems: that is, averaging on constant time slices. Some may argue that the 'natural' way of averaging for this kind of study, which involves observations, is to average the density on the null cone. It certainly is not easy to define, mainly for the reason that, as the averaging domain on the null cone is increased, the density in general will *increase* because we are looking back into the past and would thus have to account for evolution of sources. The point is that the results from such an averaging will coincide neither with those of the FLRW background nor those of this chapter, which is really just an alternative refutation of the idea that averaging reduces observations in an inhomogeneous universe to those in a FLRW universe.

The results show that it is quite easy to have areas in the inhomogeneous models which differ significantly from areas in the background, matter-averaged FLRW model. The result may either be shrinking (larger total area) or magnification (smaller total area). The presence of loops in the $\hat{R}$-z and $\hat{\rho}$-z graphs is analogous to the well known 'finger of God' effect familiar in redshift maps of the galaxy distribution.

Whilst the major aim of this chapter has been the above thesis, the choice of radial coordinate which locates the null cone will be of use in future analyses of observations in these isotropic dust models.

An important caveat is that since the LTB model does not allow for formation of caustics in the null cone of the central observer, it cannot be considered a useful model for obtaining quantitative 'real world' results. Rather this chapter should be viewed as a proof that even purely radial lensing distorts the area distance-redshift relation significantly. As will be argued in the next chapter and in the papers [212, 214], we expect caustics to skew the area towards larger values, so that most objects in the universe are demagnified.

The importance of all this is that it opens up the way for considering the effects of lensing by inhomogeneities on large angular-scale number counts and CMB observations (for example, COBE), as opposed to limiting discussion of lensing effects to small angular scales.

Chapter 7

Shrinking and its effects on cosmological area distances

7.1 Introduction

The angular - diameter distance (or 'observer area distance', equivalent up to redshift factors to the luminosity distance, see [8, 129]), lies at the heart of observational cosmology, since it is used to convert observed angles to estimated length scales and areas both at galactic distances, and on the surface of last scattering of the CMB [1].

It is usually calculated on the assumption that the universe is well represented on large scales by an exactly spatially homogeneous and isotropic FLRW geometry, presumably obtained as some kind of average of the manifestly inhomogeneous matter distribution on smaller scales [206]. However it is known this can be a bad assumption on small angular scales, because of the local inhomogeneity of matter, which causes distortion of bundles of light rays (and so gravitational lensing). Bertotti gave a power series expansion for this effect[155][2], while Dyer and Roeder derived a formula that can be used at any redshift, treating those rays that propagate in the lower density regions between inhomogeneities where shear is small [138, 142]. However this formula will not be accurate for those ray bundles that pass very close to matter, where shearing becomes important, before they reach caustics. We will argue later that it may in fact be a reasonable approximation for the average area-distance at high redshift ($z > 1$) after many caustics have occurred [3], and on angular scales which are large enough to encompass many caustics.

The usual assumption, made explicit by Weinberg [159], and accepted by most workers in the field, see for example Schneider et al [160, 142], is that although the area distance will be inaccurately represented by the FLRW formula on small angular scales due to the clumping of matter, when averaging over large enough angular scales the FLRW area distance formula

[1]This chapter was done in collaboration with G.F.R. Ellis and P.K.S. Dunsby and is based on the paper [212].

[2]Which is a consequence of the non-commutativity of smoothing the geometry and calculating null geodesics.

[3]If one assumes statistical isotropy and homogeneity of inhomogeneities, when combined with the CMB isotropy, then it is a natural assumption that the average area-distance will be isotropic too.

will be exactly correct. That this conclusion does not necessarily follow from any principle was the thrust of the previous chapter. However, there we gave no evidence as to how the averaged area distance is expected to compare to the background FLRW one.

Here we contend that areas will be quite different than in the corresponding FLRW universe on both small and large angular scales, basically because of the occurrence of caustics and associated divergence of geodesics. Apparent sizes on small angular scales will also be strongly affected. However, on larger angular scales apparent sizes will nevertheless be represented *approximately* correctly by the FLRW formula, because the caustics cause the light cone to fold in on itself.

This chapter will first give further, general arguments as to why one should expect that the focusing effects due to clumping will not average out to give the FLRW area distance. We then explain why the previous arguments either are incorrect, or do not apply to the real lumpy universe, once one follows light rays for long enough that caustics have formed in our past light cone (which is a case of considerable observational interest). Indeed 'shrinking' will occur - the distance covered on the last scattering surface for a given apparent angle in a lumpy universe, will be more than in the corresponding FLRW universe model (which is normally assumed as giving the correct geometry), so the apparent angular size of a given object can be smaller than expected if lensing is not taken into account.

While shrinking associated with any single lensing object is very small, there are a very large number of objects in the sky that will cause lensing by the time our past light cone has reached the surface of last scattering. The result of all the cumulative lensing is that the past light cone will have a fractal-like structure there. The associated observational effects are complex, and depend on angular scales. On small angular scales, for a given distance on the last scattering surface, the apparent size will be much less than in the corresponding perfectly smooth model. However due to the folding over of the light cone on itself associated with caustics, on larger angular scales the effect will average out in the sense that the observed angular sizes of large scale structures will be little affected even though the associated areas are quite different. Thus the resulting effect on particular observational relations will depend on whether it is overall angular size, or the associated observed areas, that matter in the observations.

'Shrinking' can be said to occur in any model whose are distance is different from the critical density Einstein-de Sitter (EdS) universe favoured by many theorists as a result of the inflationary universe paradigm; so in particular it occurs in a a low density FLRW universe ($\Omega \ll 1$). There the bending of space and space-time result in a larger distance being covered on the last scattering surface, for a given angular displacement at the observer, than in the corresponding EdS universe (see for example [161] for a specific calculation of this effect). Hence observational effects of shrinking due to lensing are similar in nature to those expected in a low-Ω universe relative to a critical density model. Thus one can call on analyses of the effects of negative spatial curvature to see the kinds of consequences that can follow, at least on small angular scales, from the shrinking due to gravitational lensing. As regards number counts, it is precisely this effect on areas that underlies number-count tests of q_0 (or equivalently, of Ω_0) [8, 129]. Its effect on angles underlies angular diameter estimates of Ω_0, while estimates from luminosity measurements probably involve both, because of the complex issues involved in estimating apparent magnitudes [162]. These will therefore all be affected by shrinking caused by lensing. As regards the CMB, two particular features are affected: the measured angular size of the 'Doppler peaks' in the CMB power spectrum, and the spatial power spectrum derived from observations. These are both germane to our interpretation of the CMB observations and what they tell us about the growth of structure in the universe.

The analysis of the CMB data [127, 128] generally concentrates on the scalar spectrum and in most cases assumes a perfectly smooth Einstein - de Sitter (EDS) angular - diameter distance, which neglects gravitational lensing and other observational effects which are unavoidable in a clumpy universe such as our own. The influence of gravitational lensing directly on the amplitude of the CMB temperature anisotropies has been investigated, with contradictory results, in recent years [130, 131, 134, 135, 136, 137] by calculating the change to angular functions such as the CMB correlation function. The new point raised here is the impact, due to shrinking, of gravitational lensing on the estimated *spatial* scalar power spectrum at decoupling, with the implication that similar effects might occur in the location of the Doppler peaks. We illustrate the change to the spatial power spectrum derived from observations, using the simplest model of an angular - diameter distance which allows for inhomogeneities [138, 142] as an example, after arguing why that could be an acceptable model to use in estimating shrinking effects. However we have not yet examined detailed enough models to determine over what angular scales these effects will be significant; this depends on the distribution of inhomogeneities at all redshifts between us and last scattering. Due to the occurrence of multiple lensing between us and the surface of last scattering, the angular scale below which the effect is important for angles could be as large as a few degrees, and so the effect could possibly be of significance for the analysis of the CMB Doppler peaks. It will also affect statistics of objects and number counts for sources at redshifts greater than 1 on all angular scales.

The overall aim of the chapter is to point out that shrinking due to gravitational lensing will not average out on large angular scales, in the sense that observed areas will be quite different than in FLRW geometries, and to indicate some of the possible observational consequences. Two further papers [167, 168] use different methods to confirm the reality of the effect.

7.2 Why lensing causes shrinking

In any universe, cosmological observations are dependent on the angular - diameter distance, which directly determines luminosities as well as apparent sizes [8] and indirectly determines selection effects [162] which in turn govern number counts. In an inhomogeneous universe, the area distance will differ from the corresponding relation in a smooth FLRW model [130, 138, 142][4]. However in an inhomogeneous universe which undergoes a transition to homogeneity on large scales, such as is believed to occur with our own, it is usually assumed that the angular - diameter distance will coincide exactly with that of a FLRW model when averaged over these scales [159]. If true, this depends critically on the areas of the wavefronts at the same redshift being the same in the real universe as in the FLRW model. The existence

[4] Indeed it is a theorem that if the area distance-redshift and number count-redshift relations are the same as in a Friedmann-Lemaître universe, then universe is indeed a FLRW universe.

of directions in the sky for which the wavefront is multiply sheeted due to gravitational lensing [5] shows that this cannot be assumed.

In general the relation between a scale ℓ at redshift z, and the angle θ it subtends when observed will depend on θ, on the direction of observation (represented by a unit spacelike vector $\mathbf{n}$ orthogonal to the observer's 4-velocity), and implicitly on the angular scale θ_a over which observations are averaged,[6] as well as on the redshift z of the object observed, viz:

$$\ell = r(z, \mathbf{n}, \theta_a, \theta)\, \theta\,. \tag{7.1}$$

Here $r(z, \mathbf{n}, \theta_a, \theta)$ is the angular-diameter distance in a general universe, which will be anisotropic due to the shearing effects on the ray bundle. We expect that this anisotropy will tend to zero (as in the background FLRW model) as the averaging scale increases, effectively making the angular-diameter distance isotropic in the limit of large averaging angle. Thus a fairly good approximation for the large-angle (e.g. *COBE*) experiments is that distortion (represented by large-scale shear in the null rays) is unimportant. One can, however show that small-scale shear increases the convergence of the ray bundle and thus decreases the angular-diameter distance [142, 159] for those angular scales. Larger amounts of shear cause conjugate points [170] to form in the past light cone which leads to image parity reversal and creation of multiple images due to self-intersection of the past null cone. When averaged over a large angular scale, this can result in a change in the area-distance relation.

The question, then, is how to estimate these effects when inhomogeneity causes lensing of light rays. Consider the past light cone $C^-(P)$ of the space-time event 'here and now', denoted by P. As a bundle of light rays $\mathcal{B}(d\Omega)$ generating $C^-(P)$ (and subtending a solid angle $d\Omega$ at P) pass a lensing mass L, the nearer rays are distorted in towards the central ray γ_L (with direction $\mathbf{n}$) linking P to L. Thus focusing is caused for these rays, and this can be examined by ray tracing, by use of the geodesic deviation equation, or by using the optical scalar equations. Consequently (see e.g. Figure 2 in [174], or Figure 2.3 in [142]) the area dS of the bundle of geodesics $\mathcal{B}(d\Omega)$ beyond L will be less than if L had not been there (i.e. in the reference background case, described by an exact FLRW geometry).

Now the crucial point is that we must get the overall masses right. If we take a FLRW universe and add a mass concentration to represent some inhomogeneity - a star, a galaxy, a galaxy cluster, or whatever - then the new universe has greater mass than the old; so we expect the areas to be different simply because the average mass density in a volume V of the perturbed model that includes both P and L, is different from that in the background model. We need to correct the perturbed model to get back to the original mass in this volume, so that the background model is correctly chosen to fit the perturbed model [206]. We do so by surrounding the over-density of the lump by an underdensity of equivalent mass, so that the total mass is unchanged. Or viewed differently, this is the requirement that the perturbed universe can be obtained from the background universe by rearranging masses while keeping overall mass conserved (this is the burden of the Traschen integral constraints, [175]; when they are satisfied this is equivalent to correctly fitting the background model to the lumpy universe model, [169]).

[5] Multiple imaging causes the wavefront to develop catastrophes and become multi-sheeted [188, 142].

[6] In the case of the CMB, θ_a is the resolution of the instrument. Detail smaller than this scale is lost.

Figure 7.1: A typical arc distribution behind a galaxy cluster at a given flux limit. From http://www.mpa-garching.mpg.de/Lenses/GRLens.html.

Thus when considering lensing, we must imbed the overdensity in an exactly compensating underdensity. The light rays in the underdense region will diverge more than in the background model, and those in the overdense region will converge more. The standard view [159, 142] is that these effects exactly cancel: the area in the perturbed model will be exactly the same as in the background FLRW model.

Now this does not take into account the effect of caustics on the past light cone structure and area distances. Sufficiently far down the null geodesics after passing lensing sources, caustics (and associated multiple images) will occur. The typical shape of these caustics has been presented in many papers by Roger Penrose, see e.g. [173], Figure 49; their relation to gravitational lensing is discussed for example in [142].

The point we want to make then is that after caustics have occurred, the null rays that were converging start diverging. Indeed at a caustic an infinite convergence is instantaneously converted to an infinite divergence [177]. Hence thereafter, both the rays that went through the less dense regions and those that went through a more dense region are diverging more rapidly than in the corresponding exactly smooth FLRW model. Thus the overall area (far enough down the light cone) should be greater than in the corresponding FLRW model, as all rays are subject to greater divergence.

We are interested in finding these areas after caustics have occurred. The reason is that recent Hubble Space Telescope observations have confirmed that virtually everything beyond a redshift of unity is lensed (because the entire sky is covered by intervening objects that will cause lensing), see e.g. [180, 181, 182]; and at higher and higher redshift there will be more and more lensing. This will affect all number counts at high redshift. Further, we are particularly interested in the effect this has on our past light cone by the time it has reached Σ, the surface of last scattering of the CMB, for this will influence our interpretation of the CMB data. The situation here is quite different than in relating lensing to discrete sources, for the surface of last scattering is essentially a spacelike surface; thus we are interested in the relation of the real past light cone to a spacelike surface (in contrast its relation to timelike lines, which is relevant in considering multiple lensing of discrete objects).

The key issue is, What is the area of a bundle of geodesics $\mathcal{B}(d\Omega)$ generating our past light cone $C^-(P)$ when it intersects Σ, or (almost equivalently), what is the distance ℓ traversed in this surface when one scans through an angle θ? However there is an important subtlety here.

Consider changing the direction of view $\mathbf{n}$ at P through an arc $\mathcal{A}$ as the angle of observation θ increases continuously from some arbitrary initial direction ($\theta = 0$) to a final direction ($\theta = A$), where the corresponding light rays encounter a transparent lens L centred at $\theta = \theta_L$ ($0 < \theta_L < A$), and then develop caustics before intersecting the spacelike surface Σ. As the direction at P continuously increases, the corresponding image point $p(\theta)$ in Σ will move along the image of the arc $\mathcal{A}$ (a 1-dimensional curve) in the (2-dimensional) intersection of $C^-(P)$ with Σ, resulting in a series of forward, backward, and then forward motions.

To see this clearly, we need to define the relevant light-cone parts. In the simplest case, the situation for a single lens will be as follows: the two-dimensional section of $C^-(P)$ corresponding to the arc $\mathcal{A}$ (see Figure 2 in [184], Figure 5.1 in [142], Figure 4 in [185], and Figure 25 in [186]) will develop a fold line L_1 (representing non-local intersections of the geodesics generating $C^-(p)$) and two caustic lines L_2, L_3 (representing local intersections of these geodesics), all three intersecting at the point Q which is conjugate to P along the central null geodesic γ_L. Below Q, this 2-dimensional section of $C^-(P)$ will consist of 4 parts: an outer region $C_0^-(P)$, and three others lying inside it: a side region $C_1^-(P)$ from L_1 to L_2 (intersecting Σ in the curve C_1), a back region $C_2^-(P)$ from L_2 to L_3 (intersecting Σ in the curve C_2), and a side region $C_3^-(P)$ from L_3 to L_1 (intersecting Σ in the curve C_3). The central ray γ_L together with L_1 divides $C_0^-(P)$ into left and right parts, corresponding to small θ and large θ, and intersecting Σ in the curves C_- and C_+ respectively.

For a given spacelike surface Σ beyond Q, in addition to the central ray γ_L, this lensing structure defines 4 unique geodesics through P. These are, γ_1 (with $\theta = \theta_1$), lying in the left side of $C_0^-(P)$, joining P to the intersection P_1 of L_1 and Σ; γ_2 (with $\theta = \theta_2$), lying first in the left side of $C_0^-(P)$ and then passing through L_1 (between P_1 and Q) to $C_1^-(P)$, joining P to the intersection P_2 of L_2 and Σ; γ_3 (with $\theta = \theta_3$), lying first in the right hand side of $C_0^-(P)$ and then passing through L_1 (between P_1 and Q) to $C_3^-(P)$, join P to the intersection P_3 of L_3 and Σ; and γ_4 (with $\theta = \theta_4$), lying in the right hand side of $C_0^-(P)$, again joining P to the intersection P_1 of L_1 and Σ (so P_1 is multiply imaged at P by photons arriving along γ_1 and γ_4). Before Σ, the geodesics γ_2 and γ_3 rule first $C_0^-(P)$ and then the side surfaces $C_1^-(P)$ and $C_3^-(P)$ respectively; beyond Σ (where they are tangent to the caustics L_2, L_3), they rule the back surface $C_2^-(P)$. The central geodesic γ_L, first lies in $C_0^-(P)$, and then (after passing through Q) lies in $C_2^-(P)$, where it meets Σ at P_L.

Consider now the motion in Σ of $p(\theta)$ as θ steadily increases from 0 to $A > \theta_4$. Starting at the initial point $p(0) = I$ on C_-, it moves on C_- from the left, through P_1 along C_1 to P_2, then back along C_2 to P_3, and then forward again along C_3 through P_1 onwards in C_+ to the final point $p(A) = F$ on C_+. Hence it effectively traverses the same spatial distance (between P_2 and P_3 along C_2) three times. It will be useful to calculate two distances, both for the same angular change at the observer: we need to distinguish *distance traveled* ℓ_t along the full path:

$$I \xrightarrow{C_-} P_1 \xrightarrow{C_1} P_2 \xrightarrow{C_2} P_3 \xrightarrow{C_3} P_1 \xrightarrow{C_+} F\,,$$

calculated as a line integral along that path, and *distance gained* ℓ_g - how far the image point has moved in space from its starting point, calculated by determining the shortest distance between I and F. This will be almost the same as the distance traveled along the caustic path if one omits all the closed loop segments, i.e it is essentially the line integral:

$$I \xrightarrow{C_-} P_1 \xrightarrow{C_+} F\,.$$

The difference is essentially that which occurs in a random walk - compare distance traveled by the agent (how far has his legs carried him) as against the distance moved (how far he is from where he started off). Both depend on the angle θ, but the first increases monotonically with θ, while, for each angular scale on which cusps occur, the second has a saw-tooth effect imposed on top of this uniformly increasing tendency. Because of this, the first increases with θ on average much more than the second.

The question we are interested in is how each of these distances varies quantitatively with θ, particularly for relatively large values (say $1°$ to $10°$), and how they compare with the corresponding *background distance* ℓ_b (what one would estimate as the corresponding distance traveled in the background exact FLRW geometry, where distance traveled and distance gained are the same). Correspondingly we want to compare the real area with the corresponding area in a properly fitted background FLRW universe.

It is fairly clear from this discussion that we can only tell what the resulting area is by detailed calculation, although various simple estimation methods may be employed. The change of area will be very small for any particular lens, because lensing angles are small. But the point is that the number of lensing objects is very large. Each star will cause lensing, acting as an opaque lens, as will massive planets; each sufficiently concentrated star cluster (e.g. globular clusters) will cause caustics, acting as a transparent lens, as will each galaxy and each cluster of galaxies; furthermore voids with sufficiently sharp edges will also cause lensing (they are equivalent to using the usual lensing equations with an effective negative mass density). In many cases the lensing will cause caustics to form, indeed often this will happen quite close to the lensing mass (e.g. in the case of the sun, bending of light by $1.75''$ at the limb will cause a caustic to occur in initially parallel light rays at that distance where the sun subtends an apparent size of $3.5''$ - which is .0093 parsec or .03 light years). Once a caustic occurs in our past light cone, further lensing (on moving further away down the past light cone into the past) can never remove it, but can introduce new caustics. Furthermore, each object may cause multiple cusps; for example, sufficiently far down the past light cone, an elliptic lens will cause the standard double-caustic pattern noted by various workers [188, 189].

Hence the number of caustics in our past light cone, by the time it reaches the surface of last scattering, will be extremely large, of the order of the number of stars in the observable universe, i.e. 10^{22}, and will occur in a hierarchically structured way with larger cusps (due to galaxies and clusters) superimposed on smaller cusps (due to stars and planets), leading to something like a fractal structure. It is important to realize that as we are interested here in effects on distant number counts or on the CMB spatial spectrum, rather than in detailed lensing positions related to specific sources, there is no alignment problem: *all* lensing objects that cause cusps before last scattering end up causing caustics on the LSS Σ, because it effectively occupies the entire sky; and most detectable objects will cause such cusps, because Σ is a very large distance away, corresponding to a redshift of about 1200. Thus it is likely that every point on Σ will be covered by at least a single caustic. The caustics caused on Σ by any particular lensing object L will be spherically symmetric if the lensing object is spherically symmetric, and will be centered on the null geodesic γ_L from P through L to Σ. The inner and outer caustics caused by an elliptical lens will similarly be centred on the central connecting null geodesic from the observer to the lens.

Considering this fractured structure of the real past light cone $C^{-}(P)$ by the time it hits the surface of last scattering, it is clear there are potentially significant effects on the overall area resulting from the cumulative effects of all lenses. The overall effect will remain even after averaging due to convolution of the incoming information with the detector point spread function. We will argue that distance traveled (or area distance) is substantially affected at all angular scales, but that distance gained (or angular diameter distance) is strongly affected up to some angle $\hat{\theta}_c$, but not much affected on larger angular scales. The value $\hat{\theta}_c$ depends on the clustering of matter at all redshifts up to last scattering.

7.2.1 Response to previous arguments

The paper by Weinberg [159] explicitly considers this averaging issue, and argues that there is no overall such shrinking effect. He gives two independent arguments as to why this is so. Clearly it is necessary that we answer them here. The key point is that Weinberg's main argument does not explicitly take into account the effects of caustics, which we are identifying as important.

His first argument is by explicit calculation of bending by a single finite-radius clump of matter, and the resulting intensities (based on the previous calculation by Gunn and Press [178]). However he only allows for two ray paths from the source to the observer - whereas we know that in fact in the generic case there will be three such paths. To first order in q_0 (i.e. assuming $\Omega_0 \ll 1$) he finds that the luminosity distance (estimated from the combined intensities of the two images) is the same as in the FLRW model. If we include the general third image we may expect a different result. Additionally the estimates used are only valid for $z < 1$, and do not cover the large-z case we are interested in ([178], p.400).

He then gives a second argument, based on photon conservation. This argument - essentially the same as the usual 'reciprocity theorem' effect [8] - is correct in that it determines the average number of photons intercepted by a telescope in terms of the area of a sphere drawn about the object, and works on the basis that this number is conserved (a good approximation in the context considered). The problem is that Weinberg then assumes that the area of this sphere can be calculated from the FLRW area formula, whereas this is precisely the issue that is in question. At first glance one might think the answer is obvious because here we are dealing with the up-going future light cone from the source, rather than the down-going past light cone from us; but by the reciprocity theorem, these are essentially equivalent to each other. From the viewpoint of this chapter the key point is that, just as the past light cone of the event 'here and now' will develop numerous cusps and caustics as we go further into the past from P, so will the future light cone of the source as we go further to the future from that source (provided it is far away enough in the past; and the sources we are concerned with, when dealing with the CMB, are on the surface of last scattering). Just as our past light cone develops a hierarchically structured set of caustics by the time it reaches a source S on the surface of last scattering, so the future light cone of the source S will have developed a complementary hierarchically structured set of caustics by the time it reaches us. The area of this future light cone at the present time therefore cannot be assumed to have the FLRW value; indeed this is essentially the quantity we have to calculate. Thus the area argument in Weinberg's paper does not establish the result that the averaged area distance will be the same as in a FLRW universe, as claimed; it effectively assumes this result, by assuming this area is equal to that in a FLRW model.

The third part of Weinberg's paper looks at the effect of opaque spheres, giving a formula that is a generalisation of both the Dyer-Roeder formula and the standard FLRW formula, depending on the size of the opaque spheres. This suggests the Dyer-Roeder result may be applicable for all redshifts if there is an opaque core of large enough size in the lensing object. His conclusion is 'A proper assessment of these effects requires we take into account the detailed selection procedures actually followed by observers'. Our account above did not take into account the effect of opaque centers to lensing objects; we agree that this will make the situation even more complex in the case of the smaller lenses where this is a plausible picture. This part of Weinberg's argument indicates that the actual effect will lie between that given by the Dyer-Roeder formula and the standard FLRW formula, rather than being just equal to the FLRW result; but does not explicitly take caustics into account, which will increase the 'shrinking' effect.

7.3 Observational effects

We discuss three particular effects of the shrinking effect below, after obtaining a very simple first estimate of its magnitude.

7.3.1 Simplest estimates

To estimate the relation between the various distances on the surface of last scattering Σ, we first consider the situation of a single lensing object producing a single cusp. The key issue here is what is the angular size of the cusp at last scattering, i.e. what is the angle $\theta_c = \theta_3 - \theta_2$ between the two rays that reach the outer edges P_2, P_3 of the caustic at Σ, and what is the angular separation $\theta_m = \theta_4 - \theta_1$ of the two rays that intersect Σ in the fold. This will be the upper limit to the angular separation of multiple images that this lens could produce, if we consider objects all the way back to last scattering; its value will be approximately $2\theta_c$. It will plausibly be of the order of $10''$ for galaxies and $30''$ for galaxy clusters, for we have already seen deflections or arcs on these scales, but could be larger (see below). The point then is that for that lensing object, this is the maximum deflection of light rays we have to take into account. Images will be distorted by up to this scale, but not larger. The further point is that the corresponding distances on Σ should not be calculated using the FLRW angular diameter distance, but rather a formula that allows for the overall

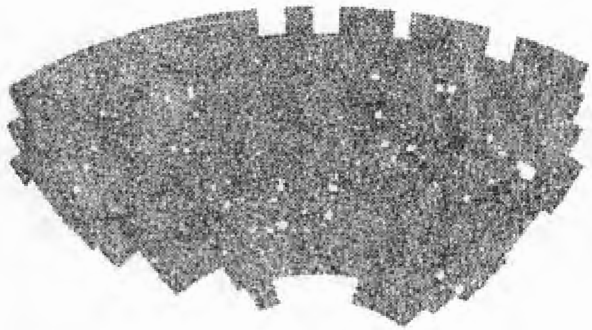

Figure 7.2: The low-redshift galaxy distribution from the APM projected survey. The large fraction of the sky covered by galaxy cores is evident even here at low redshifts. From http://www-astro.physics.ox.ac.uk/ wjs/apm-grey.gif.

shrinking effect (see next section), for these rays will be diverging more rapidly than in the corresponding background model.

Consider then a distribution of such objects, but still only taking into account single lensing (light rays only pass close enough to one such object to be appreciably deviated). Then when we consider some angular scale $A \gg \theta_c$, images on those scales will be negligibly affected by the lensing. The effect is like wrinkle glass: small scale structure is blurred but large scale structure behind is reasonably clearly visible[7]. We can immediately attain a simple estimate of the relation between the various distances mentioned above: ℓ_b will be well approximated by ℓ_g for such scales, with error at most the distance ℓ_c corresponding to the angular scale θ_c, because the distances between the widely separated rays will not be affected by more than this amount. However ℓ_t will be different: it will be approximately 3 times the distance corresponding to θ_c (calculated using the modified angular diameter distance formula), because that path will be traversed 3 times as θ increase from 0 to θ_A. This will be true for each single caustic surface encountered. For the double caustics of elliptic galaxies, there will be an increase by a factor 3 between the inner and outer caustic, and a factor 5 within the inner caustic lines. We argue below that the entire surface of last scattering will be covered by at least single caustics, so a minimum estimate of the ratio ℓ_t/ℓ_g for angles greater than θ_c is 3, and this will be approximately the same as the ratio ℓ_t/ℓ_b. The corresponding area ratio will be the square of this number, that is a minimum factor of 9. The effect is like crumpling tissue paper with numerous small-scale wrinkles, and then laying it down on a flat surface; the area it will cover on the surface is much less than its prior smoothed-out area.

Now the true situation is complicated by four factors, relative to the simple case considered so far. First, this kind of effect will occur at various angular scales, corresponding to MACHO's, stars, galaxies, and clusters of galaxies, and possibly also to star clusters, voids, and superclusters; and cusps due to these various lensing objects will probably overlap in most areas on the surface of last scattering, so there can be a large multiplicity factor m of lensing (m is the number of times the same segment of arc is traversed due to multiple cusps: 3 for a simple cusp, 5 for the interior of elliptical lenses, and so on). The area will be increased by a factor of approximately m^2 in each domain on Σ where m is constant.

[7] we thank Marco Bruni for this remark.

The first claim is that because of the multiplicity of sources occurring, which will effectively cover the entire sky for significant redshifts (see figure (7.2), the Hubble telescope deep field images and [193]), for the whole of the surface of last scattering, $m \geq 3$ (cf the diagrams in [195] for a visual impression of this multiplicity); correspondingly, the entire sky is covered by objects that will cause caustics before the surface of last scattering, but with weak lensing where θ_c is between $10''$ to $30''$ (conservative estimates for average points on the surface of last scattering, based on existing observations).

The second point is that multiple small-angle scattering can take place between P and Σ [165]. When this takes place, there are two effects: firstly, this can introduce new cusps and caustics in the past light cone structure (but cannot remove any that already exist; this is an entropy-like property, similar to what occurs in the intersection of purely gravitational idealised cosmic strings. Secondly, it will alter the angular size of existing caustics in a random way, leading to a random walk in the effective θ_c for a given lens, potentially leading to overall deflection angles that can be quite large if sufficient such scatterings take place. How large depends on the number of scatterings and angle of each one, in turn depending on the distribution of inhomogeneities all the way back to Σ.

The third factor is that large angle deflections can also occur, due to black holes. The ensemble of light rays reaching us from over the whole sky will run into numerous black holes in QSO's and galaxies, and provided they are not surrounded by material that is opaque to the CMB photons at small impact parameters (between $3M$ and $6M$) this can cause scattering by 180° and more [196]. This could lead to images being blurred on larger scales too (the effect is like bubble glass rather than ripple glass), with resulting θ_c possibly being quite large. However the further point is that although there are many of them, each is of very small size so the fraction of the sky covered by black holes is very small. Thus although this large angle scattering will happen, its integrated effect on number counts and the CMB will probably be negligible.

The final factor is that we must remember that the lensing we are interested in is taking place in a curved space-time. The lensing equation [142] is defined in terms of the following quantities: D_{ds}, the angular diameter distance from lens to source, D_s, the angular diameter distance from observer to source, D_d, the angular diameter distance from lens to observer; β, the unperturbed (background) angle from source to observer relative to optical axis from source to lens, θ, the apparent position of the source, represented by the angle of the image relative to the optical axis from source to lens, $\hat{\alpha}$, the actual deviation of light ray at lens, and α, the apparent deviation of light ray as seen by the observer. Then the basic relation is:

$$\beta = \theta - \frac{D_{ds}}{D_s}\hat{\alpha}(\xi) \tag{7.2}$$

((2.15a) in [142]), where $\xi = \theta D_d$ is the impact parameter at the lens. Now the fundamental point is that in a cosmological context, in general D_s and D_{ds} will reach a maximum at a redshift z_* (1.25 in the case of a critical density FLRW universe, about 4 for a low density FLRW model). Thus *distant objects can cause large apparent deflections if D_s is small*; and for sources at redshifts greater than z_*, the further away they are, the smaller D_s is. Combined with multiple scattering, this could result in apparent deflections of a degree or more, even from quite ordinary astronomical objects. The detail depends on what formula we use for D_s (discussed in the next section). This complements the discussion in chapter 5 on the monotonic decrease of Σ_{crit} with source redshift for a single lensing object. To understand how Σ_{crit} varies with many lenses, is one of the outstanding challenges in gravitational lensing.

It is clear then that in a realistic model of the universe, the past light cone is an extremely complex object covered with cusps on many angular scales. The issue is what fraction of sky is covered by objects causing scattering on various angular scales, leading to a distribution of probability for the whole sky of deflection angles θ_c, and so a most-likely deflection $\hat{\theta}_c$ on average over the sky; and what is the multiplicity m from superposition of all these objects at an average point on the surface of last scattering Σ. Our view is that the probability of θ_c will be large for all angles from microarcseconds to at least $30''$, and probably extending up to several minutes or even degrees because of multiple scatterings combined with the effect of a curved space-time and taking into account the effect of superclusters. There will be a tail up to larger angular scales due to black holes, but of very low amplitude. Correspondingly, an average point on Σ will have multiplicity $m \geq 3$, leading to an area shrinking factor, when averaged on large scales (or over the whole sky), of at least 9, and plausibly 25 or greater.

7.3.2 Broad nature of observational effects

The effect on observations will be appreciable once the cumulative effect of lensing has started to build up - say at a redshift of 1 and beyond.

In measurements that depend on area effects, the change due to shrinking will broadly correspond to the squares of the multiplicity. The actual observational effects will depend on the 2-dimensional distribution of cusps on surfaces of constant redshift such as Σ, which cannot easily be estimated from the 1-dimensional projections considered here. As emphasized above, the effect from each single lens is small; but at the distances we are considering the *entire sky* is lensed, indeed is multiply lensed on different angular scales. Number counts will be altered because the areas covered by the light rays in a given solid angle are larger than estimated from the FLRW formula. It is clear that the effect could be significant, particularly for number counts at high redshift, or at lower redshift but based on surveys in small angular diameter regions. This effect will occur at large as well as small angular scales.

The angular measurements of CMB fluctuations will also be affected. However the effect is quite complex. The distance traveled along the surface of last scattering Σ by the measuring beam is the distance traversed ℓ_t (cf above), which is greater than the distance gained ℓ_g (the extent of each cusp is traversed three times, rather than once). The area corresponding to distance traversed should be used in defining area distance, for it is the effective area for number counts and also for the last scattering surface from which light is received within a beam of width θ. Distance gained is relevant for comparing observations of large scale spatial features with estimates of their physical size.

If we consider an observer sweeping a narrow beam across the sky and measuring incoming radiation in that direction, at the surface of last scattering this beam will traverse the cusps that occur in the intersection of the past light cone with the surface of last scattering, consequently moving forward, backward, and then forward each time such a cusp occurs [see Section 2 above] and almost performing a random walk when one takes into account the whole hierarchical structure of these cusps . Thus any particular small-scale temperature fluctuation will be sampled several times as it is scanned both forwards and backwards by the measuring beam; hence any Gaussian fluctuations on these scales will be measured as non-Gaussian (in effect, the actual spatial distribution is convolved with the saw-tooth pattern). For large scale inhomogeneities, this multiple sampling of a given spatial pattern will make little difference to the observed shape of the inhomogeneity, for it will occur on small angular sizes only; small changes in amplitude will occur on those scales, and will be smoothed out by the instrument response function. What is measured on large scales is determined by the distance gained ℓ_g, which tells us when the sampling point reaches new large-scale features of the inhomogeneous distribution of matter on the last scattering surface. The smaller backward and forward traverses are then averaged over in an effective coarse-graining. The corresponding shrinking factor relative to the background will be close to unity on scales larger than the peak in the distribution of θ_c over the sky. To determine this distribution requires detailed modeling.

7.4 Estimates obtained using the Dyer-Roeder distance

With this geometric situation in mind and for illustrative purposes, we neglect the shear [138] as a first approximation to attain an analytic formula for rough estimates of the basic shrinking effect. It does not allow for folding over: thus this estimate will apply to small angular scales, in particular to estimating the sizes of the caustic surfaces, and when squared will give an estimate of the area effect at all angular scales.

In some directions at some small angular scales, this neglect of shear is unjustified. However this approximation - corresponding to being in a lower density universe in the gaps between galaxies or clusters - should be fairly accurate for a narrow beam for a large part of its traverse through space-time. When it passes close to inhomogeneities, as explained above it will focus more than in the lower density regions but will then develop caustics and start diverging. Probably this divergence will more than make up for the previous convergence when a large distance has been traversed, so the overall effect will be to give even greater shrinking than in the gaps between galaxies and other inhomogeneities. We make the simplifying assumption that on average the matter moving through inhomogeneities, rather than the spaces between them, at a significant distance beyond formation of cusps suffer the same diverging effect as if they had only traversed the lower density regions. This assumption needs checking; we believe it is conservative, in that it will underestimate shrinking. It receives some support from Weinberg's calculation for opaque spheres (an effect we have not explicitly taken into account in this chapter). In any case it gives a simple estimate of the area effect against which other calculations can be compared.

The observed angular - diameter distance in the absence of shear becomes locally isotropic and can be characterized by the redshift only, so $r = r(z)$, where we will deal with the simpler conceptual case of an ideal experiment with perfect resolution, $\theta_r, \theta \to 0$ since the complications of averaging are non-trivial. Our simplifying assumption implies it does not matter how many caustics occur in any particular direction; but this is probably not true. One may guess that in fact the more caustics there are and so the more overlap there is, the larger the shrinking will be. However this needs confirmation by detailed calculation. As emphasized, in this section we do not take into account the folding over of the light cone onto itself, so this calculation represents a first estimate of the area shrinking on all scales and the angular shrinking on scales less than $\hat{\theta}_c$.

In this chapter for simplicity we consider the magnitude of the shrinking effect relative to a EDS model, so we follow Dyer and Roeder (1973) and Schneider *et al.* (1992) [138, 142] by introducing a smoothness parameter α [8] such that the matter in the universe is described by two components; a smooth dust background of average density $\alpha\Omega_d$ and a proportion $[1-\alpha]\Omega_d$ in compact clumps, which under the assumption of negligible shear, do not affect the angular - diameter distance. We then define the ratio:

$$\gamma(\alpha, \Omega, z) = \frac{r_{\alpha,\Omega}(z)}{r_{1,1}(z)}\,, \tag{7.3}$$

where $r_{\alpha,\Omega}(z)$ is the angular - diameter distance for a universe with smoothness parameter $\alpha(z)$ and density parameter Ω at redshift$z$, (so $r_{1,1}(z)$ is the angular - diameter distance for a pure FLRW model with density parameter $\Omega = 1$). In particular we note that in the FLRW case the following relation exists between arbitrary Ω and $\Omega = 1$ universes [142]:

$$\begin{aligned} \gamma(1, \Omega, z) &= \frac{\Omega z - (2-\Omega)(\sqrt{\Omega z + 1} - 1)}{\Omega^2(z + 1 - \sqrt{z+1})} \\ &> 1 \quad \text{if} \quad \Omega < 1\,. \end{aligned} \tag{7.4}$$

Thus we see that shrinking relative to the EDS model occurs even in pure FLRW models – a well-known result – but is enhanced by including inhomogeneities.

From equations (7.3) and (7.4) we see that a linear scale ℓ on the surface of last scattering, subtending an angle θ_{EdS} in a smooth EDS universe, actually subtends an angle:

$$\theta = \frac{1}{\gamma(\alpha, \Omega, z)}\theta_{EdS}\,. \tag{7.5}$$

provided this equation applies at that scale. For example the angular size of the Hubble radius at decoupling, usually quoted to be at an angle of about $1°$, actually subtends an angle $\sim \frac{1°}{\gamma}$ where γ defines the shrinking factor at those angular scales for the redshift $z_{SLS} \approx 1000$, corresponding to the surface of last scattering. It is possible that this factor is significant. The issue is whether this angular scale is larger or smaller than $\hat{\theta}_c$; we argue above that it is possible that it is larger.

We include for generality a radiation energy density term Ω_r, which may be important in low Ω universes which are radiation dominated at decoupling. Then the equation for the angular - diameter distance generalizing the Dyer-Roeder distance to that case is obtained from the transport equations for the optical scalars, on neglecting the shear:

$$P(z)\ddot{r} + Q(z)\dot{r} + \left[\frac{3}{2}\Omega_d\alpha + 2(1+z)\Omega_r\right]r = 0\,, \tag{7.6}$$

$$P(z) = (1+z)(1 + \Omega_d z + z(2+z)\Omega_r)\,, \tag{7.7}$$

[8] not to be confused with the deflection angle mentioned above; this will not occur again.

$$Q(z) = (3 + \frac{\Omega_d}{2} + \frac{7\Omega_d z}{2} + \Omega_r + 8\Omega_r z + 4\Omega a_r z^2)\,. \tag{7.8}$$

The corresponding initial conditions are [170]:

$$r(z=0) = 0\,, \qquad \dot{r}(z=0) = 1\,. \tag{7.9}$$

This equation gives distances in units of $\frac{c}{H_0}$ and reduces to the usual Dyer - Roeder equation when $\Omega_r = 0$ [9]. If in addition $\alpha(z) = \alpha_0$, a constant, this equation can be converted into the hypergeometric equation [138] or the Legendre differential equation [170], so that solutions exist for all α_0 and Ω. Finally Linder (1988) has obtained the solutions for a large number of different limiting cases [130].

For the numerical work [10], we choose a smoothness parameter of the form:

$$\alpha(z) = 1 - (1 - M_*)\left[\frac{1-\alpha_5}{1-M_*}\right]^{\frac{z}{5}}\,, \tag{7.10}$$

where $M_* \equiv \alpha(0)$ is the present proportion of mass in smoothly distributed form and $\alpha_5 \equiv \alpha(5)$ is the same quantity at a redshift of 5, which corresponds to the present outer limit of observed quasars. We parametrize α_5 by an evolution index, $p, (p > 1)$, which is constant for each simulation, so that:

$$\alpha_5 \equiv \alpha(z=5) = \frac{M_* + p - 1}{p}\,, \tag{7.11}$$

with more rapid evolution into clumps simulated by using smaller p. The neutral baryonic component in the observable universe is almost entirely in clumpy form today and even at $z = 5$ there appears to be very little neutral gas in the intergalactic medium [146] (using the standard interpretation of the Gunn - Peterson test).

The two parameters (M_*, p) are more difficult to estimate in the case of dark matter, which is a partial restatement of the bias problem. Only a fairly small proportion of the dark matter is expected to be in compact form today since the inhomogeneities are still approximately linear on scales larger than about $10h^{-1}Mpc$. In particular it should be noted that $\alpha(z)$ depends on the scale one is averaging over to obtain the "clumps", with the limiting behaviour that $\alpha(z) \to 1$ when averaged over the whole sky. Thus the shrinking factor is strictly speaking dependent on observational scale even when shear is neglected, making the impact on the estimated power spectrum even more intricate.

Because of this uncertainty, and for generality, we plot our results for a wide range of M_* (figure 7.3; see also table 1). For $\Omega = 1$ we calculate the range of γ – factors as a function of $M_* \equiv \alpha(z = 0)$, the smoothness parameter today. We take as the smallest reasonable γ – factor, the value $\gamma_{min} = 1.08$.

There is an interesting question which is also of practical importance: what is the asymptotic behaviour $(z \to \infty)$ of $\gamma(\alpha, \Omega, z)$? This is important for stability reasons since the surface of last scattering is ill - defined and has finite thickness, with redshift estimates

[9] The equation has the form of a general Mathieu equation with parametric "frequency" and "damping" dependencies. Note that the ansatz $r(z) = exp(-\frac{1}{2}\int \frac{Q(\eta)}{P(\eta)}d\eta)\chi(z)$ will eliminate the $\dot{r}$ term.

[10] We integrated eq.(7.6) using a Runge - Kutta - Merson method, which is $\mathcal{O}(h^5)$ and gives an estimate of the local error.

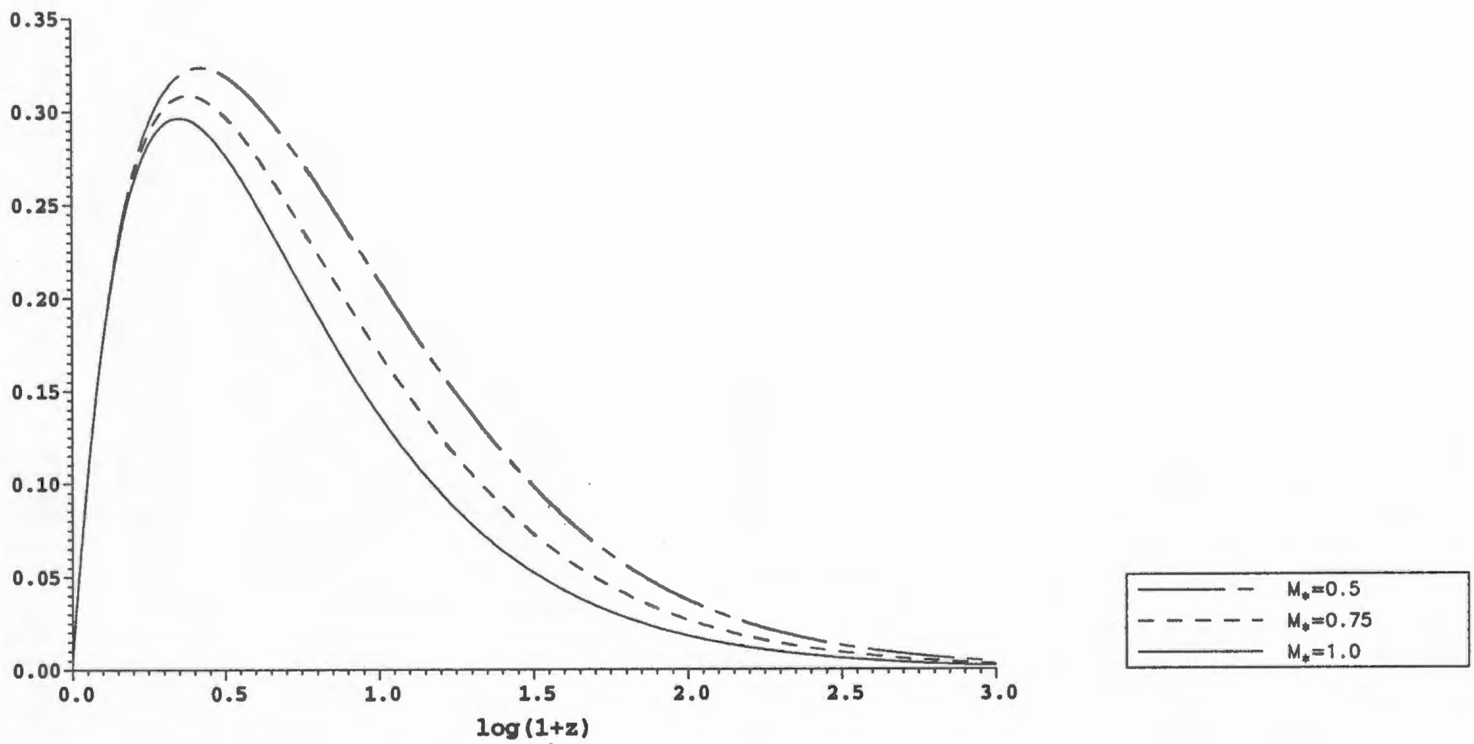

Figure 7.3: The area-distance as a function of the redshift and present smoothness parameter, M_*, the proportion of the energy density in smoothly distributed form. The solid line corresponds to the background FLRW model with $M_* = 1$ and no shrinking.

M_*	p	γ	γ^{6+n} $(n = 1.5)$	γ^{6+n} $(n = 1)$
0.95	2.0	1.08	1.78	1.71
0.9	1.5	1.26	5.66	5.04
0.8	2	1.40	12.47	10.54

Table 7.1: Shrinking factors, γ, for various evolutionary parameters (n is the spectral index). See below for discussion of M_*, p.

ranging between $z_{SLS} = 1000 - 1500$. Further it is found that $\gamma(z)$ monotonically increases with redshift. This result implies that our estimates of the shrinking effect will be slightly increased if we assume the surface of last scattering lies at $z_{SLS} = 1500$ instead of 1000. For very large redshifts the shrinking factor may approach a non- zero constant. It is easy to show that a constant γ is a solution of the governing evolution equation in the asymptotic region $z \to \infty$ and we conjecture that this constant should be determined by the low-redshift behaviour of $\alpha(z)$, since asymptotically $\alpha(z) \to 1$ for generic perturbed FLRW models.

7.5 The new power spectrum

As all inhomogeneities on the surface of last scattering appear smaller than they would in a FLRW model, for angles less than $\hat{\theta}_c$ (hence the term "shrinking" [132]), we expect the emphasis in the power spectrum, in the domain affected by angular shrinking, to shift to larger wavelengths (smaller k) due to the scaling properties of the Fourier transform.

In general the observed CMB will be a patchwork of regions with different shrinking factors, since we observe through underdensities and overdensities, so that a circular region of CMB anisotropy will be seen to be dilated uniformly only if it is small and has an angular size close to the resolution of the instrument or if it is large enough to enclose a large number of caustics. The CMB patterns on medium scales will have different parts deformed by different amounts, depending primarily on how many overlapping caustics of different angular scales occur in any particular region, leading to a highly non-conformal mapping from the surface of last scattering (or the spatial hypersurface the anisotropies were produced in) onto the celestial sphere. We consider the simplest case: how the power spectrum changes when we include the effects of a constant shrinking factor. By scaling fluctuations on the surface of last scattering by a factor $\gamma \equiv \gamma(\alpha, \Omega, z = 1000)$ and then taking its Fourier transform, how is the new power spectrum $|\hat{\delta}_k|^2 \equiv |\hat{\delta}(k)|^2$ related to the old one, $|\delta(k)|^2$? Consider the scaling of the random density contrast field $S(\mathbf{x})$:

$$S(\mathbf{x}) \equiv \frac{\rho(\mathbf{x}) - \bar{\rho}}{\bar{\rho}} \quad \mapsto \quad S(\frac{\mathbf{x}}{\gamma}) \tag{7.12}$$

Here $\rho(\mathbf{x})$ is the density at decoupling, $\bar{\rho}$ is the average density and $\gamma \equiv \gamma(\alpha, \Omega, z)$. Note that scaling all inhomogeneity sizes by γ (active transformation) is equivalent to reducing the coordinate $\mathbf{x}$ by γ to $\frac{\mathbf{x}}{\gamma}$ (passive transformation), as done above.

The Fourier transform of $S(\mathbf{x})$ in $3-$dimensional space is: [11]

$$\delta(k) \equiv \delta_k = \int \frac{d^3\mathbf{x}}{(2\pi)^3} e^{i\mathbf{k}\cdot\mathbf{x}} S(\mathbf{x}) \tag{7.13}$$

and the new spectrum

$$\hat{\delta}(k) = \int \frac{d^3\mathbf{x}}{(2\pi)^3} e^{i\mathbf{k}\cdot\mathbf{x}} S(\frac{\mathbf{x}}{\gamma}) \tag{7.14}$$

which after relabeling of the spatial variable and some rearrangement gives:

$$\hat{\delta}(k) = \gamma^3 \delta(k\gamma) \tag{7.15}$$

In the case of a power-law spectrum which is expected on large and medium scales at decoupling, shrinking changes our estimation of it by increasing the amplitude via:

$$|\delta(k)|^2 = Ak^n \Rightarrow |\hat{\delta}(k)|^2 = (A\gamma^{6+n})k^n \,, \tag{7.16}$$

where A is the normalization constant. The spectral index is unchanged [12] in the case of a power-law spectrum (since power laws are scale invariant). Thus in the power-law regime the true amplitude of the spectrum at decoupling is larger by the factor γ^{6+n} than that found from observations assuming an EDS model.

Considering only the effects of gravitation and free streaming [ignoring photon diffusion, Sakharov oscillations and the integrated Sachs-Wolfe effect [133, 147]] on an initially power law spectrum, we find that if shrinking is uniform across all the relevant scales, it increases with k to a maximum and then turns around and decreases for $k > k_{eq} \approx \frac{\Omega^2}{2}$, where k_{eq} is the wave number corresponding to the Hubble scale in Mpc^{-1} at the change over from radiation to matter domination. The theoretical functional form of the power spectrum for $k \gg k_{eq}$ is predicted to be [148]:

$$|\delta(k)|^2 = B\frac{\ln^2(k/k_{FS})}{k^3} \qquad k \gg k_{eq} \,, \tag{7.17}$$

where B is a normalization constant determined by matching to the spectrum at large scales and k_{FS} describes the smallest perturbations which are not washed out due to the free-streaming of the particles forming the dark matter. If this is the true form of the spectrum at small scales, then we will observe something different due to the shrinking effects of the inhomogeneities.

The power spectrum at small scales, *assuming* it takes the form given in equation (7.17), is obtained by "inverting" the process followed at large-scales, since we now wish to predict

[11] Strictly speaking $\delta(k)$ should depend on $\mathbf{k}$, but because of the assumed isotropy (ergodicity [211]) of the density perturbations, we impose the restriction that it depend only on k. This assumption is one of simplicity imposed on the primordial matter perturbations (Gaussianity), and will have to be reconsidered when caustics and the effects of shear are included on estimates of the spectrum because of the anisotropy of the angular-diameter distance in that case.

[12] The invariance of n under shrinking is only true in the approximations made in this chapter. n will be a function of γ when the scale-dependence of shrinking due to shear and averaging is included.

the observed spectrum, and not the true spectrum. By using our formula for the new spectrum, equation (7.15), one obtains:

$$\begin{aligned} |\hat{\delta}(k)|^2 &= B\gamma^{-6+3}k^{-3}\left[\ln(k/(\gamma k_{FS})^2\right] \\ &= \gamma^{-6+3}\left\{|\delta(k)|^2 - \frac{B}{k^3}\ln\gamma\left[\ln(\frac{k^2}{(k_{FS}^2\gamma)})\right]\right\}, \end{aligned} \tag{7.18}$$

and hence (for $\gamma > 1$) $|\hat{\delta}(k)|^2 < |\delta(k)|^2$ for this k-range, which shows that shrinking will lead to observations which over-estimate the true spectrum and hence lies below the spectrum derived from observations ignoring shrinking. This may be important for CMB experiments at scales less than a degree such as the OVRO [149] and South Pole 89 [150] experiments.

If one solves the perturbed kinetic theory problem for the coupled photon- baryon fluid before decoupling one gets the Doppler peaks, or Sakharov oscillations. The possible effects of lensing on the position of these Doppler peaks is important, since it has been proposed [151] that the position of the first Doppler peak can be used to determine the value of Ω. Since our results show that for angular scales less than $\hat{\theta}_c$, the angular power spectrum is shifted relative to the one estimated without considering lensing, it is possible shrinking will change the position of the Doppler peaks. In particular, the position of first Doppler peak is determined by the angular size of the Hubble scale at decoupling, which is determined by the geometry and dependent on γ. Hence, unless lensing effects are included, the value of Ω estimated using small-angle experiments in the future, could be wrong. Detailed work including the effects of caustics should be included when examining the very-small scale experiments planned for the future.

7.6 Conclusions

In this chapter we have asked the question of how lensing due to inhomogeneity will change estimates of number counts and of spatial functions at decoupling, in particular the mass fluctuation power spectrum.

Number counts at high redshift will definitely be affected, because they depend on areas rather than apparent positions; and areas overall, even when averaged over large angular scales, will be strongly affected (by at least a factor 9). The precise factor depends on the clumping history of the universe. One estimate is the square of the multiplicity factor due to caustics (Section 3), another is the square of the angular shrinking factor estimated from the Dyer-Roeder equation (Section 4). Detailed simulations will be necessary to determine the best estimate, which will depend strongly on redshift.

The relation between CMB observations and theory may be significantly affected (Section 5). It will affect the observed fluctuations at small scales, where the spectrum may not be power-law. A possible further application is to the Doppler peak distribution in the spectrum at small-scales. It has been suggested that the position of the first Doppler peak is sensitive to the value of Ω, but insensitive to most other physical variables, and hence could be a test of the geometry of the universe. Here we point out that shrinking might shift the position of the Doppler peaks.

It is important to answer the question of how shrinking differs from other investigations of lensing of the CMB, which have invariably concentrated on the change to the *angular* correlation function. The standard method uses perturbed geodesics on a flat background, calculating the perturbation effects up the unperturbed null cone. However, to calculate the change to spatial functions at a given redshift, as required for comparison with structure formation theories, one cannot use smooth background formulae. One must recalculate the effect on converting angles to distances when going down the null cone, by including the effects of inhomogeneity.

As with previous investigations on the effects of inhomogeneity and gravitational lensing on the propagation of light, our results will depend fairly sensitively on the amount of matter in non-linear, clumpy form and on its evolutionary history [130, 204]. We also emphasize that the main approximation made in the numerical estimates, namely the neglect of the shear, may be unacceptable. More sophisticated analysis is required to determine what influence shear has, the impact of caustics and the scale- dependence of shrinking, but the results obtained from this conceptually simple effect, highlight some of the pitfalls that abound in making use of standard models. The same reason makes analysis of the implications for structure formation and inflationary models worthwhile.

We have emphasized the effects of shrinking due to the formation of caustics, but in fact it will also occur when weak lensing takes place, as can be seen from detailed analyses. Here we aim to show the effect is possible and indeed probable. The paper [168] confirms that it does occur in a particular case where we can obtain exact solutions of the Einstein equations, while [167] shows the effect is generically one of shrinking rather than magnification, as is plausible from the discussion given here. This conclusion is supported by studies, e.g. [199], showing that most sources are demagnified rather than amplified when lensing occurs and caustics are taken into account. This will strongly affect the selection effects that underlie number counts at high redshift.

Chapter 8

The error function and the Kink Soliton

8.1 Introduction

In this, the final chapter of the thesis, we digress slightly from of cosmology to applied mathematics and the issue of the integration of the Gaussian. The production of Gaussian random fields is crucial in modern cosmology, particularly when discussing the CMB and structure formation theories. Almost all modern cosmological tests are based on statistical indicators, such as correlation functions, which all try to probe the nature of the underlying parent distribution. When it comes to comparing theory with experiment, one must be able to generate simulations with statistics described by a given distribution. The most common of these is the Gaussian distribution, both because of its ease of use and the central limit theorem. To generate Gaussian random fields, requires either directly or indirectly, the error function, the definite integral of the Gaussian, which is the subject of this chapter [1].

Now there is an inherent asymmetry between integration and differentiation which makes integration somewhat of an art form, and which is perhaps best exemplified by the lack of an elementary indefinite integral of the celebrated Gaussian:

$$\int \exp\left(\frac{-(x-\beta)^2}{\sigma^2}\right) dx \tag{8.1}$$

The fact that such an integral does not in fact exist follows from the work of Laplace [230, 231]. However, the Gaussian integral is fundamental, finding applications in statistics, error theory and many branches of physics. In fact, anywhere one has Gaussian distributions, cummulatives of these distributions will involve the above integral. Only special case definite integrals of e^{-x^2} are known, the most famous being:

$$\int_0^\infty e^{-x^2/\sigma^2} dx = \frac{\sqrt{\pi}\sigma}{2} \tag{8.2}$$

[1] Based on the paper [229].

In addition there is the series expansion [232]:

$$\int_0^x e^{-u^2} du = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{(k-1)!(2k-1)} x^{2k-1} \tag{8.3}$$

Now in practise one can evaluate the integral accurately by numerical methods or tables, but in many cases it would be preferable to have an analytical solution, even if it were not exact, as long as the maximum error were very small and the approximation were simple [2].

It turns out that there exists a function well known in the analysis of nonlinear partial differential equations whose derivative is very close to Gaussian - the kink soliton:

$$\phi(x) \equiv A \tanh(bx - c\beta) \tag{8.4}$$

with derivative:

$$\chi(x) \equiv Ab\left(1 - \tanh^2(bx - c\beta)\right) \tag{8.5}$$

where $A, b, c,$ and β are all real constants. The graphs of e^{-x^2} and $\chi(x)$ are shown in figure (1). [3] The kink soliton is the positive, time-independent, topological solution to the non-linear $1+1$ dimensional partial differential equation:

$$\phi_{tt} - \phi_{xx} = 2b^2(\phi - \frac{1}{A^2}\phi^3) \tag{8.6}$$

where a subscript denotes partial derivative with respect to that variable. The solution to this equation is topological because the boundary conditions at $x = \pm\infty$ are different. Leaving the physical origin of ϕ behind, it is interesting to examine the series expansion of $\tanh(x)$:

$$\tanh(x) = \sum_{k=1}^{\infty} \frac{2^{2k}(2^{2k}-1)}{(2k)!} B_{2k} x^{2k-1}\,, \quad \text{valid for } x < \frac{\pi}{2} \tag{8.7}$$

which should be compared with eq. (8.3) for $\int e^{-x^2} dx$. Here B_k are the Bernoulli numbers with generating function $t/(e^t - 1)$. We see that although the coefficients differ in each case, the powers of x in the expansions are identical. Further both $\chi(x)$ and e^{-x^2} have the property that their derivatives can be re-expressed in terms of themselves and $\phi(x)$ or powers of x respectively. These observations shed some light on the foundations of the approximation.

8.2 Details of the approximation

Turning to practical issues, we are left with choosing the constants, A, b, c to optimise the approximation of eq. (8.1). We need three constraints to fix the three parameters. First we require that the Gaussian and $\chi(x)$ have the same symmetry axis. This requires the argument of tanh to vanish at $x_* = \beta$ which immediately implies from eq. (8.5) that $c = b$.

[2] Several rational function approximations exist but they are rather complicated [233].

[3] One can consider a one-parameter family of approximations to the Gaussian given by replacing $x \to x^\epsilon$ in eq. (8.5) which give better fits when $\epsilon \neq 1$, but which do not have indefinite integrals as far as is known to the author.

At this stage we have a choice, dependent on whether we are interested in an approximate solution for small or large x. For large x, a constraint is obviously that our new approximation, $\phi(x)$, must give *exactly* the same result as eq. (8.2) when differenced at infinity and the origin. This will ensure convergence of our approximation. Since $\tanh(x) \to 1$ as $x \to \infty$, and $\tanh(0) = 0$, this implies from eq. (8.4) that:

$$A = \frac{\sqrt{\pi}\sigma}{2}$$

Finally we can impose that $\chi(x) = e^{-(x-\beta)^2/\sigma^2}$ at some point, i.e, we match the derivatives. We will choose $x = \beta$ as the simplest. This gives:

$$Ab = 1 \Longrightarrow b = \frac{2}{\sqrt{\pi}\sigma}$$

In fact the two are equal at another point as can be seen from figure (1). Our analytical approximation, which is very accurate for large x, is therefore:

$$\phi(x) = \frac{\sqrt{\pi}\sigma}{2} \tanh\left(\frac{2}{\sqrt{\pi}\sigma}(x-\beta)\right) \simeq \int e^{-(x-\beta)^2/\sigma^2} dx \tag{8.8}$$

where in this paper $\simeq$ is understood as meaning asymptotic convergence, as $x \to \infty$ and bounded error $\forall x$. From figures (1,2) we see that the kink derivative underestimates the Gaussian at small $(x-\beta)^2$ and overestimates it at large $(x-\beta)^2$.

Alternatively if one is interested in $\int_0^u e^{-x^2/\sigma^2} dx$ where $u \leq 4\sigma$ say, then this will not be good enough, since the error in our approximation is strongly confined to small x. Instead we can impose that $\phi(x)$ must give the exact result, not at infinity, but at the end of the interval, i.e. at u. Thus we impose:

$$A \tanh(b(u-\beta)) = \int_0^u e^{-(x-\beta)^2/\sigma^2} dx \tag{8.9}$$

In addition we need to match the derivatives $\chi(x_*) = e^{-(x_*-\beta)^2/\sigma^2)}$ at some point x_* as before, and then solve the equations for A, b. It is an open question which matching point yields the best results. For illustrative purposes we choose $x = \beta$ and again find $A = 1/b$, so that substituting in eq.(8.9) gives us a nonlinear root-finding problem for A. The right-hand side can be found for example, from tables of the error function, erf(x). This yields an approximation which is exact at $x = u$ and hence a much better approximation for small x, but which is invalid for $x \gg u$. The extension to cases with variable lower limit of integration is obvious and will not be considered.

One might be tempted to generalise eq. (8.4) to a one-parameter family of approximations to the error function:

$$\Delta_p(x) = A \tanh^p(bx) \tag{8.10}$$

which have derivative:

$$\Delta'_p(x) = Abp \tanh^{p-1}(bx) \mathrm{sech}^2(bx) \tag{8.11}$$

However, since for $p \neq 1$, $\Delta'_p(0) = 0$, they are not really suitable as approximations to a Gaussian. Rather they are skewed distributions with maxima at $x > 0$. It turns out however, that they will be useful later.

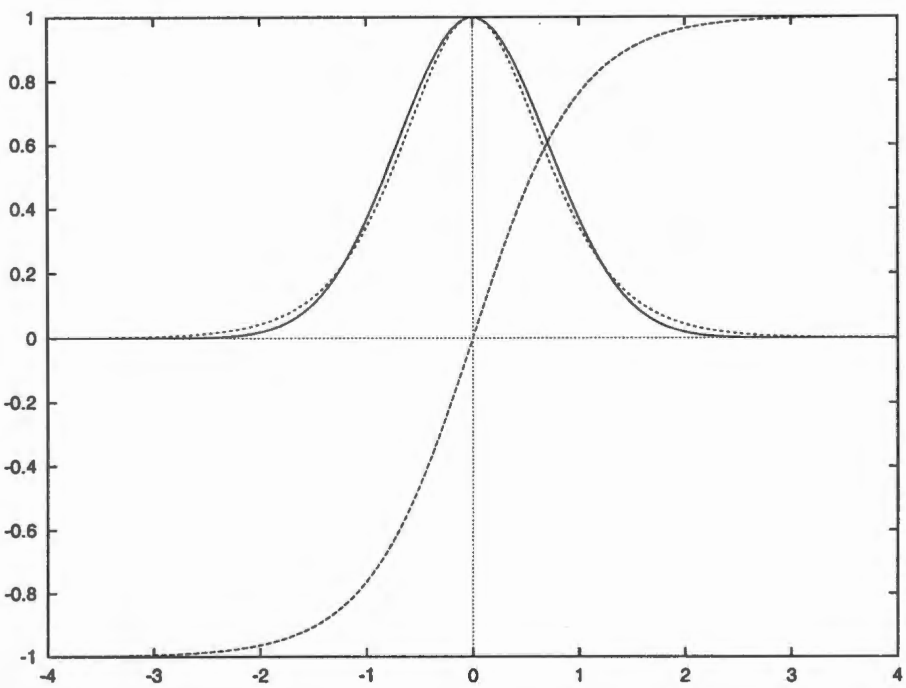

Figure 8.1: Plot of e^{-x^2} (solid line), $\chi(x)$ (dotted line) and $\tanh(x)$ (dashed line), which is the kink soliton.

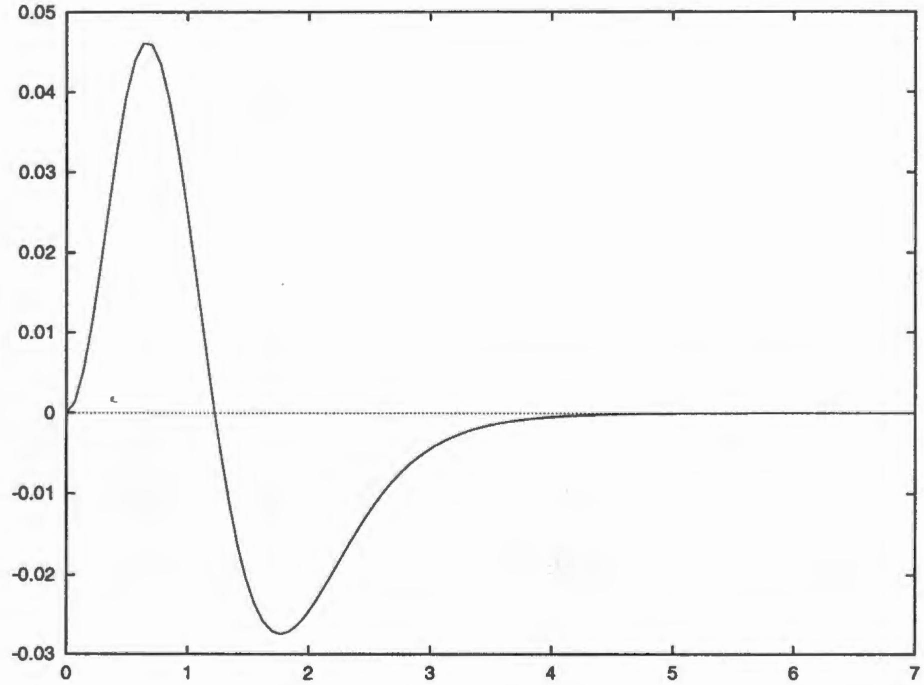

Figure 8.2: Plot of the difference between e^{-x^2} and $\chi(x)$.

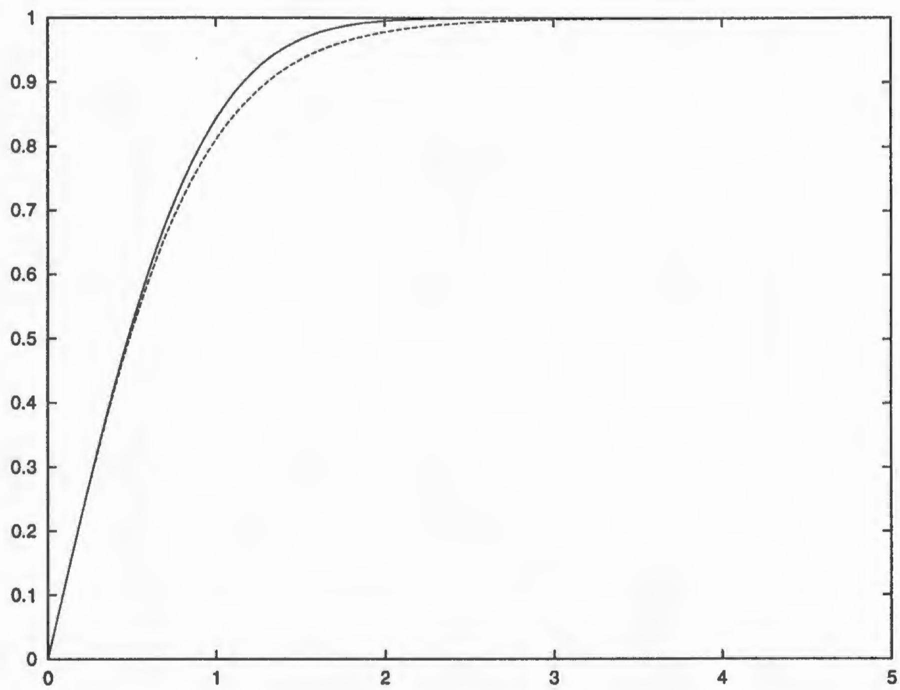

Figure 8.3: Plot of the error function, and the soliton approximant, $\phi(x)$. The maximum difference occurs at $x = 1.12$ and is 3.91%. The error drops below 1% for $x \geq 2.3$ and converges exponentially to zero.

For testing our approximation we will use the $\phi(x)$ valid for large x, denoted $\phi(x)_L$, given by eq. (8.8). The crucial question is of course, how good is this approximation ? It turns out that it is very good in most cases, as can be seen from figures (3) and (4). The maximum error from using $\phi(x)_L$ is 3.91% at $x = 1.12$. However as discussed earlier, if one is interested in the result for small x, and x_1 is small, then this is not the best approximation to use. In practise, the error drops off very quickly due to the exponential nature of $\tanh(x)$. For example, the error in estimating $\mathrm{erf}(x)$ drops below 1% for $x \geq 2.3$ and at $x = 5$ the error is 2.51×10^{-5}. The error as a function of x is plotted in figure (4).

8.3 Improving the approximation

The shape of figure (4) is, in fact, rather startling because of its simplicity. From the graph it has a single local maximum and two points where the concavity changes. Hence although it cannot be written down explicitly in terms of elementary functions [231], it can be approximated very closely. Several fitting shapes were tried, such as the log-normal and Poisson distributions, but the best was found to be a generalised Maxwell-distribution:

$$E(x) = \alpha_1 x^n \exp(-\frac{x^2}{\alpha_2}) \tag{8.12}$$

For the case used in the figures, that of $\mathrm{erf}(x)$, the best parameters for reducing the maximum error (i.e. minimising w.r.t. the sup-norm $||\cdot||_\infty$) were (see figure (5)):

$$\alpha_1 = 0.062\ , \quad n = 2.27\ , \quad \alpha_2 = 1.43 \tag{8.13}$$

which reduced the *maximum* error to 0.15%. It is also likely that our choice of function and parameters for $E(x)$ is not optimal, since formal optimisation was not used, but was based rather on a numerical investigation of the parameter space $\{\alpha_1, \alpha_2, n\}$.

Further, since the required $E(x)$ is a skewed Gaussian with maximum at non-zero x we can profitably employ the functions given by eq. (8.11), originally introduced to model the Gaussian, as fits for the error. In this case our approximation becomes:

$$\int_0^x e^{u^2}\,du = \frac{\sqrt{\pi}}{2}\left[\tanh(\frac{2}{\sqrt{\pi}}x) + (\alpha_3 \tanh^p(x))'\right] \tag{8.14}$$

where $'$ denotes derivative w.r.t. x. For $\alpha_3 = 0.23$ and $p = 9.7$ the error is at most 9×10^{-3}. By suitable generalisation of the second term it is possible to increase the accuracy to the level of the generalised Maxwell distribution, but for simplicity and because of its suggestiveness, we leave it in the above form.

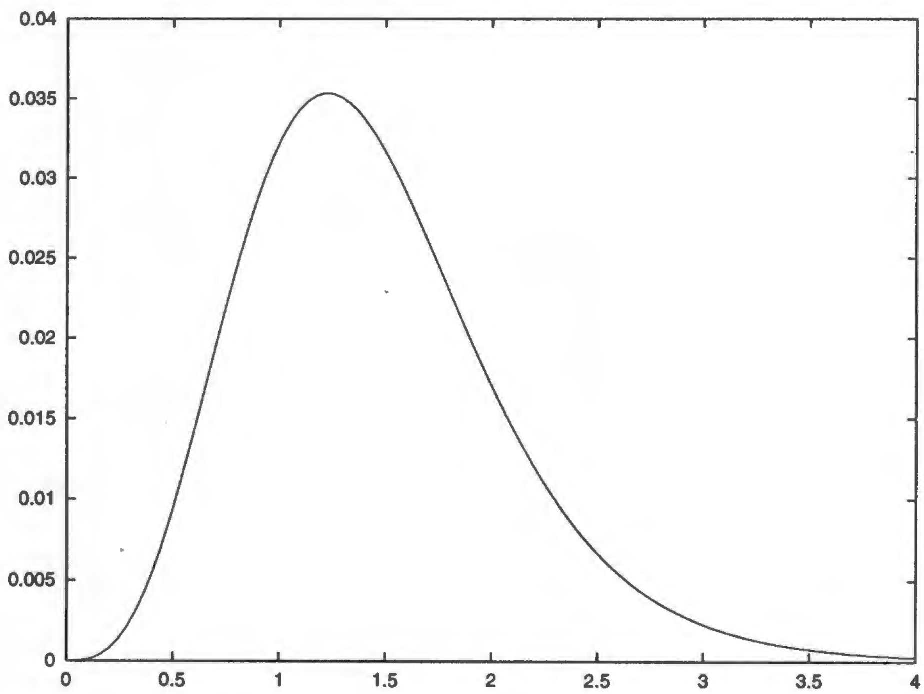

Figure 8.4: The difference of erf(x) and $\phi(x)_L$. This is closely approximated by log-normal distributions or generalised Maxwellians of the form $\alpha_1 x^n e^{-x^2/\alpha_2}$.

In the case of the error function we have explicitly that ($\beta = 0$):

$$\text{erf}(x) \simeq \tanh(\frac{2}{\sqrt{\pi}}x) + E(x) \tag{8.15}$$

where $\text{erf}(x) \equiv \Phi(x) = 2/\sqrt{\pi}\int_0^x e^{-u^2}du$ is the error function. Similarly the complementary error function is given by: $\text{erfc}(x) = 1 - \tanh(\frac{2}{\sqrt{\pi}}x) - E(x)$.

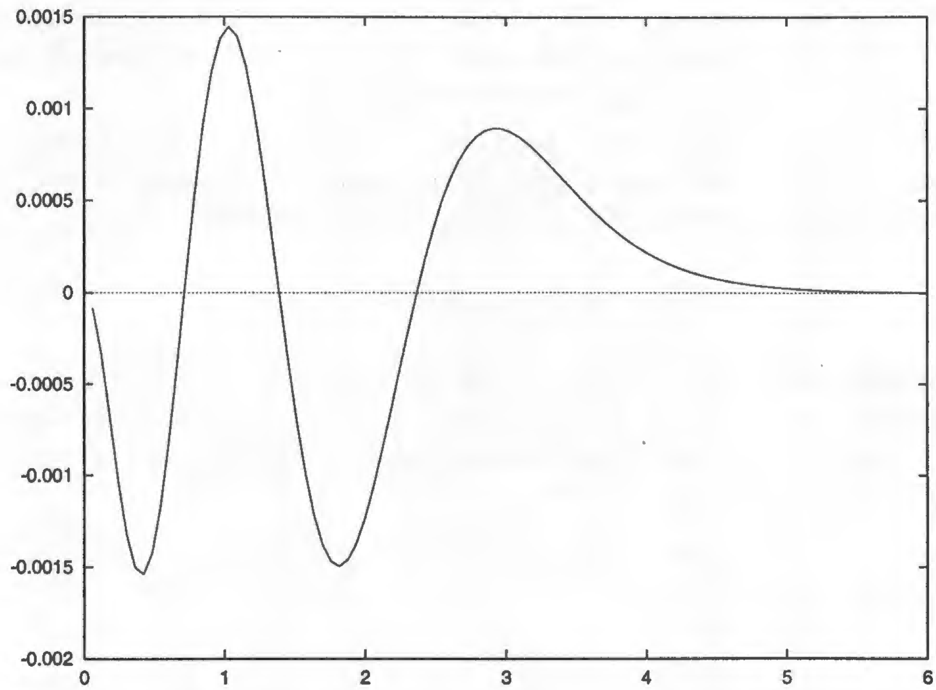

Figure 8.5: The final error after modeling figure (4) by the generalised Maxwell distribution of eq.(8.12). The maximum error is about 0.15%.

8.4 Moments of the soliton

A fundamental feature of a Gaussian distributed random variable is that all moments above the second, such as the skewness, are zero. From this it follows that the sum of error distributed random variables is itself error distributed. A natural question to ask is how well the soliton approximation preserves this feature.

To make this more precise: given the distribution $P(x)$, we may define the partition function $Z(J)$ [4] via:

$$Z(J) = \int P(x) e^{Jx} dx \tag{8.16}$$

From the "free energy" $F(J) = \ln Z(J)$ we may now define the n-th moment, M_n, of $P(x)$ as:

$$M_n \equiv \frac{d^n}{dJ^n} F(J)|_{J=0} \tag{8.17}$$

Thus in the case of a Gaussian distribution with zero mean, it is easy to show that the free energy is a quadratic function of J. Hence the only non-zero moment is the second, i.e. the variance, as claimed above. In the case of the soliton approximant we have:

$$Z(J) = \int [1 - \tanh^2(2x/\sqrt{\pi}\sigma)] e^{Jx} dx \tag{8.18}$$

[4] We use the notation $Z(J)$ because of its ubiquitous use in statistical physics. In the case where x is a function, $Z(J)$ becomes a path-integral and derivative becomes functional derivative in eq. (8.17).

which is unfortunately not known analytically, so we resort to numerical analysis. Using the Gaussian case as a testbed we approximated the free energy with an 8-th degree polynomial:

$$F(J) \sim \Sigma_{n=0}^{8} \alpha_n J^n \qquad (8.19)$$

For a Gaussian $\alpha_n = 0, \quad n \geq 3$. Using a least-squares method, the error, i.e. the largest α_n coefficient which is zero in the exact case but non-zero in the fit, was $\alpha_3 = 2.535 \times 10^{-8}$. Each subsequent coefficient was roughly an order of magnitude smaller than the preceding one.

In the case of the soliton approximation, given by eq. (8.5), the error was 2.309×10^{-3} again for the cubic term, and again with roughly $\alpha_{n+1} \sim \alpha_n/10$.

α_3	α_4	α_5
-2.535×10^{-8}	4.202×10^{-9}	-4.002×10^{-10}
-2.309×10^{-3}	4.308×10^{-4}	-4.389×10^{-5}
α_6	α_7	α_8
2.182×10^{-11}	-6.322×10^{-13}	7.532×10^{-15}
2.586×10^{-6}	-8.144×10^{-8}	1.071×10^{-9}

Table (1) shows a comparison between the coefficients of the free-energy polynomials for the terms higher than cubic for the exact Gaussian and the soliton approximant $\chi(x)$. An interesting thing to note is that, although the accuracy is at the level one might expect, i.e. $\sim 10^{-3}$, the pattern of the terms is identical; namely both the signs and the decrease in the coefficients have the same behaviour in both cases. This suggests that numerical errors will be "coherent", i.e. the errors one has from numerical integration of the Gaussian will be of the same nature as those one obtains from the soliton approximation. This is perhaps obvious given the similarity of their power series (see eq.s (8.3,8.7)) but will not be true for other approximants in terms of e.g. rational functions [233].

We leave this discussion by noting that inclusion of $E(x)$, via e.g. eq. (8.12), in the calculation of moments will reduce the above errors considerably, presumably by a factor of at least 10^2.

8.5 Applications

Let us now consider a small sample of applications. A primary example is in the theory of statistics. If we have a uniformly distributed random variable χ and we desire a random variable y with statistics given by a distribution f, first define the integral $F(x) = \int_0^x f(\chi) d\chi$. Then $y = F^{-1}(x)$ will have the same distribution as f, where F^{-1} denotes the inverse of F, on the interval $[F^{-1}(0), F^{-1}(x)]$.

In particular if, as is often the case, we want to generate a realisation of a Gaussian random distribution, $f = \exp(-x^2/\sigma^2)$, then with our approximation, $F(x) = \phi(x)$ (we have dropped the error correction term $E(x)$ for simplicity) and the inverse $\phi^{-1}(x)$, gives us our random variable. In this case if $y = \phi(x)$, then:

$$\phi^{-1}(x) = \frac{\sqrt{\pi}\sigma}{2}\tanh^{-1}(\frac{2}{\sqrt{\pi}\sigma}x) \tag{8.20}$$

which has the same form as $\phi(x)$ with the replacement $\tanh \rightarrow \tanh^{-1}$ so that both the integral and inverse are essentially trivial. This avoids the necessity of using traditional Monte Carlo methods to calculate Gaussian distributions.

A related problem occurs in the study of structure formation from gravitational collapse from Gaussian initial conditions, a standard assumption. The Press-Schecter formalism [234], gives the cumulative mass function $f(> M)$, which is the number of objects (such as galaxies) with mass greater than M:

$$f(\geq M) = 1 - \mathrm{erfc}\left(\frac{\delta_c}{\sqrt{2}\sigma(M,z)}\right) \tag{8.21}$$

where $\delta_c, z \in R$ and σ is the variance of the distribution. This can be estimated immediately using eq. (8.15).

One place where error functions are ubiquitous is in diffusion theory, since the decaying Gaussian is a solution to the standard diffusion equation. In the case where there is an extended distribution of diffusing material, situated at $x < 0$ for example, the solution is instead given by:

$$C(x,t) = \frac{C_0}{2}\mathrm{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) \tag{8.22}$$

where D is the diffusion constant. Indeed the error function appears any time there is a summation of the effects of a series of line sources each of which has an exponential distribution, both in finite and infinite media, as discussed in great detail in [235].

Further, the error function can be related to special values of the degenerate hypergeometric function, ${}_1F_1(\alpha;\gamma;z)$. In particular:

$${}_1F_1(\frac{1}{2};\frac{3}{2};-x^2) \simeq \frac{\sqrt{\pi}}{2x}\tanh(\frac{2}{\sqrt{\pi}}x)$$

Our final example comes from the theory of parabolic cylinder functions, $D_p(z)$, which are solutions to the differential equation:

$$\frac{d^2u}{dz^2} + (p + \frac{1}{2} - \frac{z^2}{4})u = 0 \tag{8.23}$$

with $u = D_p(z)$ and for integer values of $p = n$, they are related to the Hermite polynomials, $H_n(z)$ by $D_n(z) = 2^{-n/2}e^{-z^2/4}H_n(\frac{z}{\sqrt{2}})$. Finally we may write, for the special cases of $n = -1, -2$:

$$D_{-1}(z) \simeq e^{z^2/4}\sqrt{\frac{\pi}{2}}\left[1 - \tanh(\sqrt{\frac{2}{\pi}}z)\right] \tag{8.24}$$

$$D_{-2}(z) \simeq -e^{z^2/4}\sqrt{\frac{\pi}{2}}\left[\sqrt{\frac{2}{\pi}}e^{-z^2/2} - z(1 - \tanh(\sqrt{\frac{2}{\pi}}z))\right] \tag{8.25}$$

$$\tag{8.26}$$

8.6 Conclusions

In this chapter we have presented a function approximating erf(x) to better than 4% $\forall\ x$, with exponential convergence as $x \to \infty$. This solution is simply the kink soliton, $\phi(x) = \tanh(2x/\sqrt{\pi})$ and can be optimised for accuracy if the error function at small values of the argument is required.

Further we have found a solution with maximum error of 0.15% by adding a generalised Maxwell distribution to the kink soliton, equations (8.12), (8.14). Future work should be aimed at finding truly optimal solutions. Finally a few applications were discussed, particularly to diffusion dynamics and to the generation of Gaussian random fields.

Bibliography

[1] A. Linde, *Particle Physics and Inflationary Cosmology*, (Harwood, Chur. Switzerland, 1990)

[2] R. H. Brandenberger, Rev. Mod. Phys., **57**, 1, (1985)

[3] L.H. Ford, in "Cosmology and Gravitation", Proc. VII Brazillian School, Ed. M. Novello, Editions Frontiers, (1993)

[4] C. Bernard, Phys. Rev. D**9**, 3313, (1974)

[5] A. Ashtekar, *Polymer geometry at Planck scale and quantum Einstein equations*, in proc. of GR-14, (1996)

[6] B.A. Bassett and V. Villalba, *in preparation*, (1997)

[7] D.W. Jordan and P. Smith, *Nonlinear Ordinary Differential Equations*, Clarendon Press, Oxford, (1987)

[8] G.F.R. Ellis, *Relativistic Cosmology, Proc. of the Int. School of Physics "Enrico Fermi"*, ed R. K. Sachs, (New York: Academic Press), (1971)

[9] R. Rajaraman, *Solitons and Instantons*, Elsevier, Amsterdam (1982)

[10] P. Morse and H. Feshbach, *Methods of Mathematical Physics*, McGraw-Hill, New York (1953)

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

[12] J.J. Halliwell and S.W. Hawking, Phys. Rev. D **31**, 1777 (1985)

[13] M. Bruni, G. F. R. Ellis, and P. K. S. Dunsby, Class. Quantum Grav. **9**, 921 (1992)

[14] S. Anderegg and V. F. Mukhanov, Phys. Lett B, **331**, 30 (1994)

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

[16] V. F. Mukhanov, H. A. Feldman and R. H. Brandenberger, Phys. Rep. **215**, 203 (1992)

[17] E. R. Harrison Rev. Mod. Phys., **39**, 862 (1967)

[18] J.M. Stewart and M. Walker, Proc. Roy. Soc. Lond., A **341**, 49 (1974)

[19] G.F.R. Ellis and M. Bruni, Phys. Rev. D **40**, 1804 (1989)

[20] G.F.R. Ellis, MNRAS, **243**, 509 (1990)

[21] A.H. Guth and S. Pi, Phys. Rev. D**32**, 1899 (1985)

[22] D. Boyanovsky, H.J. de Vega and R. Holman, *To appear in the Proceedings of the "2nd Journée Cosmologie", Observatoire de Paris, Eds. H.J. de Vega and N. Sánchez, World Scientific*

[23] E. Calzetta and B. L. Hu, Phys.Rev. D **52**, 6770 (1995)

[24] D. Boyanovsky, H.J. de Vega, R. Holman and J.F.J. Salgado, preprint LPTHE-96/32, hep-ph/9608205 (1996)

[25] M. Novello and J.M. Salim, in *Galaxies and Cosmology*, Vol. II of Handbook of Astronomy, Astrophysics and Geophysics, Eds. V.M. Canuto and B.G. Elmegreen, Gordon and Breach Science Publishers (1988)

[26] L. Kofman, A. Linde, and A. Starobinsky, Phys. Rev. Lett. **73**, 3195 (1994)

[27] E.W. Kolb and M.S. Turner, *The Early Universe*, (Princeton, 1990)

[28] H. Kodama & T. Hamazaki, Prog.Theor.Phys. **96**, 949 (1996)

[29] Y. Nambu, and A. Taruya, Preprint, gr-qc/9609029 (1996)

[30] N. D. Birrell and P. C. W. Davies, *Quantum Fields in Curved Space* (Cambridge University Press, Cambridge, 1982)

[31] T. S. Bunch and P.C.W. Davies, Proc. R. Soc. Lond., **A360**, 117 (1978)

[32] S. W. Hawking, Ap. J, **145**, 551 (1966)

[33] L. P. Grishchuk, Phys. Rev. Lett. **70** 2371 (1993)

[34] M. Novello, J. Salim, M. C. Motta, S. E. Jorás, and R. Klippert Preprint CBPF-NF-009/94

[35] M. Bruni, P. K. S. Dunsby, and G. F. R. Ellis, Ap. J. **395** 34 (1992)

[36] G. F. R. Ellis, M. Bruni, and J. Hwang, Phys. Rev. D. **42** 1035 (1990)

[37] G. Grignani, P. Sodano, C. A. Scrucca, J.Phys. A **29**, 3179 (1996)

[38] N. Deruelle, V. F. Mukhanov, Phys.Rev. D **52** 5549 (1995)

[39] P. K. S. Dunsby, M. Bruni and G. F. R. Ellis, Ap. J. **395** 54 (1992)

[40] L.P. Grishchuk, Phys. Rev. D **50**, 7154 (1994)

[41] J.M. Stewart, CQG, **7**, 1169 (1990)

[42] M. Yoshimura, Prog. Theo. Phys. **94**, 873 (1995)

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

[44] W. Hu and N. Sugiyama, Submitted to Phys. Rev. D, astro-ph/9411008 (1994)

[45] J. Polchinski, *String Duality*, Preprint NSF-ITP-96-60, hep-th/9607050, (1996)

[46] M. Novello and J.M. Salim, in *Galaxies and Cosmology, eds. V.M. Canuto and B.G. Elmegreen*, Gordon and Breach Publishers, (1988)

[47] G.F.R. Ellis and P. Hogan, *to appear* Class. Quant. Grav. (1997)

[48] G.F.R. Ellis and P.K.S. Dunsby, *submitted to* Ap. J, astro-ph/9410001 (1994)

[49] W. M. Lesame, G. F. R. Ellis, P. K. S. Dunsby, Phys.Rev. D **53**, 738 (1996)

[50] P.K.S. Dunsby, B.A. Bassett and G.F.R. Ellis, *to appear* Clas. Quant. Grav. (1997)

[51] R. Maartens, Phys.Rev. D **55**, 463 (1997)

[52] S. Mattarese, O. Pantano and D. Saez, Phys. Rev. D **47**, 1311 (1993)

[53] M. Bruni, S. Mattarese and O. Pantano, Ap. J, **445**, 958 (1995)

[54] W.M. Lesame, P.K.S. Dunsby and G.F.R. Ellis, Phys. Rev. D **52**, 3406 (1995)

[55] A. Barnes and R.R. Rowlingson, Class. Quantum Gravity, **6**, 949, (1989)

[56] S.W. Goode, Phys. Rev. D **39**, 2882 (1989)

[57] P. Szekeres, J. Math. Phys., **6**, 1387 (1965)

[58] M. P. Ryan, Ann. Phys. (NY), **65**, 506 (1971)

[59] N. J. Cornish, J. J. Levin, Phys.Rev.Lett. **78**, 998 (1997)

[60] B. K. Berger, D. Garfinkle, E. Strasser, Class.Quant.Grav. **14**, L29 (1997)

[61] D.N. Vollick, Phys. Rev. D **48**, 3585 (1993)

[62] R. Penrose, GRG, **7**, 31 (1976)

[63] R. S. Ward, Phys. Lett., **A 61**, 81 (1977)

[64] B. K. Berger, Preprints, astro-ph/9512003, astro-ph/9512004

[65] J.D. Barrow and F. Tipler, Phys. Rep. **56**, 372 (1979)

[66] B. K. Berger, and V. Moncrief, Phys. Rev. D **48**, 4676 (1993)

[67] J. Isenberg, V. Moncrief, Ann. Phys. (N.Y.), **199**, 84 (1990)

[68] B. A. Bassett, in *Proc. of Dark Matter 1996*, eds. P. Salucci and M. Persic, (1996)

[69] T. Buchert, MPG-Preprint, *submitted to* MNRAS, (1992)

[70] G.F.R. Ellis, Pvt. communication. (1994)

[71] Dunsby, P.K.S. pvt. communication (1994)

[72] E. Bertschinger and A.J.S. Hamilton, Ap.J. **435**, 1 (1994)

[73] G.F.R. Ellis and P.K.S. Dunsby, *submitted to Ap. J*, astro-ph/9410001, (1994)

[74] W. Hu, D. Scott, & J. Silk, CfPA-TH-94-12, astro-ph/9402045 (1994)

[75] J. Hwang, Phys. Rev. D **48**, 3557 (1993)

[76] P.K.S. Dunsby Phd. Thesis, (QMW, London Univ.) (1993)

[77] G.F.R. Ellis, C. Hellaby and D.R. Matravers, Ap. J, **364**, 400 (1990)

[78] J. Wambsganss, R. Cen, J. P. Ostriker, *submitted to Ap. J*, astro-ph/9610096 (1996)

[79] E. Bertschinger and B. Jaim, Ap.J. **431**, 495 (1994)

[80] H. van Elst *et al*, *submitted to Clas. Quant. Grav.*, gr-qc/9611002 (1997)

[81] W. B. Bonnor, Class. Quantum Grav. **12**, No 2, 499, February (1995)

[82] W. B. Bonnor, Class. Quantum Grav. **12** No 6, June (1995)

[83] R. van de Weygaert, and A. Babul, Ap.J, **425**, L59 (1994)

[84] R. van de Weygaert, in proceedings "Mapping, Measuring and Modelling the Universe", Valencia 1995, eds. P. Coles & V. Martinez

[85] J. Ibáñez, A & A, **124**, 175 (1983)

[86] H. Mutoh, T. Hirai, K. Maeda, Preprint astro-ph/9608183 (1996)

[87] R. Cen, and J.P. Ostriker, Ap.J **399**, L113 (1992)

[88] V. Antonuccio-Delogu, and S. Colafranceso, Ap.J, **427**, 72 (1994)

[89] J. Silk, and R.F.G. Wise, Phys. Rep. **231**, 293 (1993)

[90] R.G. Carlberg, Ap.J, **367**, 385 (1991)

[91] P. Coles and F. Lucchin, *Cosmology*, (Wiley: 1995)

[92] A. de Oliveira-Costa, G.F. Smoot and A.A. Starobinsky, astro- ph/9510109 1995

[93] W. Hu and N. Sugiyama, Ap.J, **436**, 456, (1994)

[94] W. Hu and N. Sugiyama, *submitted to* Ap.J Lett, (1995)

[95] R.K. Sachs, & A.M. Wolfe, ApJ, **147**, 73 (1967)

[96] P.J.E. Peebles and J.T. Yu, Ap. J, **162**, 815 (1970)

[97] P.J.E. Peebles, Ap.J, **153**, 1 (1968)

[98] C.P. Ma and E. Bertschinger, MIT preprint, astro-ph/9506072, (1995)

[99] M. Tegmark, astro-ph/9511148, Proc. Enrico Fermi, Course CXXXII, Varenna, (1995)

[100] W. Hu, PhD Thesis UC Berkeley, astro-ph/9508126 (1995)

[101] A. Albrecht, D. Coulson, P. Ferreira and J. Magueijo, to be published in Phys. Rev. Lett., (1996)

[102] H. Kodama and M. Sasaki, Int. J. Mod. Phys., **A1**, 265, (1986)

[103] M. Kamionkowski, D. N. Spergel, Ap.J, **432**, 7 (1994)

[104] W. Hu, N. Sugiyama, Phys. Rev. D**51**, 2599 (1995)

[105] M. A. Janssen *et al*, astro-ph/9602009, *submitted to* Ap.J (1996)

[106] A. Loeb, A. Kosowsky, Ap.J. **469**, 1 (1996)

[107] G. F. R. Ellis and P. K. S. Dunsby, UCT Preprint (1997)

[108] M. Tegmark and G. Efstathiou, MNRAS, **281**, 1297 (1995)

[109] Joao Magueijo, Andreas Albrecht, Pedro Ferreira, and David Coulson, Phys.Rev. D**54**, 3727 (1996)

[110] A. Albrecht *et al*, Phys. Rev. Lett. **76**, 1413 (1996)

[111] H. Russ, M. Soffel, C. Xu and P.K.S. Dunsby, Phys. Rev. D, **48**, 4552, (1993)

[112] U. Seljak and M. Zaldarriaga, astro-ph/9603033 (1996)

[113] J.R. Bond, B.J. Carr & C.J. Hogan, AP.J, **367**, 420, (1991)

[114] A. Kogut in *Current Topics in Astrofundamental Physics*, ed. N. Sanchez and A. Zichichi, World Scientific, (1992)

[115] G.F. Smoot, UCB Preprint, astro-ph/9505139 (1995)

[116] J.G. Bartlett & A. Stebbins, Ap J, **371**, 8, (1991)

[117] A.S. Kompaneets, Sov. Phys. - JETP **4**, 730, (1957)

[118] B. Jones, in *VII Brazilian School of Cosmology and Gravitation*, ed. M.Novello, 1994, Editions Frontieres

[119] M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, (1959)

[120] R.G. Crittenden, D. Coulson, N.G. Turok, Phys.Rev. D **52** 5402 (1995)

[121] N. Kaiser, MNRAS, **202**, 1169 (1983)

[122] M. Zaldarriaga & D.D. Harari, Phys. Rev. D**52** 3276 (1995)

[123] M. J. Rees, Quart. J.R.A.S., **28**, 197 (1987)

[124] J. P. Vallee, Ap.J, **360**, 1 (1990)

[125] B. A. Bassett, *Questionae Mathematicae*, **19**, 417 (1996)

[126] U. Seljak, *Preprint*, astro-ph/9505109, (1995)

[127] G. F. Smoot *et al.*, *Ap. J. Lett.* **396**, L1 (1992).

[128] E. L. Wright *et al.*, *Ap. J. Lett.* **396**, L13 (1992).

[129] S. Weinberg, *Gravitation and Cosmology*, (Wiley and Sons:New York, 1973)

[130] E. V. Linder, *Astro. Astrophys.* **206**, 199 (1988).

[131] E. V. Linder, *MNRAS*, **243**, 353

[132] E. V. Linder, *Astron. Astrophys.* **206**, 190 (1988).

[133] H. Kodama & M. Sasaki, *Int. J. of Mod. Phys.*, **1**, 265 (1986).

[134] M. Sasaki, *MNRAS* **228**, 653 (1989).

[135] S. Cole and G. Efstathiou, *MNRAS* **239**, 195 (1989).

[136] K. Tomita and K. Watanabe, *Prog. Theor. Phys.* **82**, 3, 563 (1989).

[137] L. Cayon *et al*, Ap. J **403**, 471, (1993)

[138] C. C. Dyer and R. C. Roeder, *Ap. J. Lett.* **180**, L31 (1973).

[139] R. Maartens, G. F. R. Ellis, W. R. Stoeger, Phys. Rev. D **51**, 1525 (1995)

[140] S. Bildhauer and T. Futamase, GRG, **23**, 1251, (1989).

[141] S. Bildhauer and T. Futamase, MNRAS, **249**, 126, (1991)

[142] P. Schneider, J. Ehlers and E. E. Falco *Gravitational Lenses*, (Springer-Verlag, Berlin, 1992).

[143] S. Seitz and P. Schneider, Max- Planck-Institut preprint MPA 775, *Submitted to Astro. Astophys.* (1993).

[144] B. A. C. C. Bassett and P. K. S. Dunsby, UCT preprint-94/5, (1994).

[145] M. Wilson, Ap.J, **273**, 2, (1983).

[146] R. Bar-Kana, Preprint astro-ph/9401050 *submitted to Phys.Rev* D (1994).

[147] W. Hu, D. Scott and J. Silk, Preprint astro-ph/9402045 *Submitted to Ap.J. Lett.* (1994).

[148] T. Padmanabhan, *Structure Formation in the Universe*, Cambridge University Press, (1993).

[149] A. C. S. Readhead *et al* *Ap.J* **346** 556 (1989).

[150] P. Meinhold & P. Lubin, *Ap.J. Lett.* **370**, 11, (1991).

[151] M. Kamionkowski, D.N. Spergel, and N. Sugiyama, Ap.J Lett, **426**, 57, (1994).

[152] P. Coles and G.F.R. Ellis, *Nature*, **370**, 609 (1994).

[153] P.J.E. Peebles, Principles of Physical Cosmology (Princeton: Princeton University Press), (1993).

[154] X. Luo, Univ. of California Preprint, astro-ph/931200, (1993).

[155] B. Bertotti Proc. Roy. Soc. Lond. **A294**, 195 (1966).

[156] G. F. R. Ellis and W. R. Stoeger, *Classical Quant. Grav.* 4, 1697 (1987).

[157] G. F. R. Ellis, in *General Relativity and Gravitation*, Ed. B. Bertotti et al (Reidel, 1984), 215-288.

[158] C. C. Dyer and R. C. Roeder, *Ap. J. Lett.* **180**, L31 (1973).

[159] S. Weinberg, *Ap. J*, **208**, L1, (1976).

[160] J. Ehlers and P. Schneider, *Astron Astrophys* **168**, 57 (1986).

[161] G. F. R. Ellis and R. Tavakol: *Class Qu Grav* **11**, 675-688 (1994).

[162] G. F. R. Ellis, J. J. Perry and A. Sievers: *Astronom Journ* **89**, 1124-1154 (1984).

[163] G. Jungman, M. Kamionski, A. Kosowsky, and D.N. Spergel. *Phys Rev* **D54**, 1332 (1996).

[164] W. Hu and M. White. *Astrophys JOurn* **471**, 30 (1996).

[165] T. Fukushige, J. Makino, and T. Ebisuzaki *ApJ* **436**, L107 (1994).

[166] J. A. Munoz and M. Portilla. *Astrophys Journ* **465**, 562 (1996).

[167] N. Mustapha, B. A. Bassett, C. W. Hellaby and G. F. R. Ellis, UCT preprint, *submitted* to *Class Quant Grav*, (1996).

[168] D. Solomon and G. F. R. Ellis: UCT Preprint (1996).

[169] U. Seljak, astro-ph/9505109 (1995)

[170] S. Seitz and P. Schneider, Max-Planck-Institut Prepritn MPA 775; (1993)

[171] Y. Kakagi, T. Okamar, and T. Fukuyama. *Int Journ Mod Phys* **D** 4, 685 (1995)

[172] W. Hasse, M. Kriele, and V. Perlick. *Class Quant Grav* **13**, 1161 (1996)

[173] R. Penrose. *Techniques of Differential Topology In Relativity.* Society for Industrial and Applied Maths (Philadelphia, 1972).

[174] S. Refsdal, *Mon Not Roy Ast Soc* **128**, 23 (1964).

[175] J. Traschen, *Phys Rev* D **29**, 1563 (1984); **D31**, 283 (1985).

[176] G. F. R. Ellis and M. Jaklitsch: *Astrophys. Journ.* **346**, 601-606 (1989).

[177] S. Seitz, P. Schneider, and J. Ehlers, Class. Quant. Grav. **11**, 2345 (1995)

[178] W. H. Press and J. E. Gunn, *Ap J* **185**, 397 (1973).

[179] S. D. J. Gwyn and F. D. A. Hartwick. *Astrophys Journmn* **468**, l77 (1996)

[180] I. Smail, R. S. Ellis and M. J. Fitchett, *MNRAS* **270**, 245 (1994).

[181] I. Smail, R. S. Ellis, M. J. Fitchett and A. C. Edge, *MNRAS* **273**, 277 (1995)

[182] I. Smail, W. J. Couch, R. S. Ellis and R. M. Sharples, *Ap J* **440**, 501 (1995)

[183] D. W. Hogg, R. Blandford, A. Kundic, C. D. Fassnacht, and S. Malhotra *Astrophys Journ* **467**, L73 (1996).

[184] R. Nityanda. In Gravitational lensing, ed. Y Mellier, B Fort and G Soucail, Springer Lecture Notes in Physics Volume 360 (Springer, 1990).

[185] B. Fort and Y. Mellier, *Astron Astrophys Rev* **5**: 239 (1994)

[186] S. Refsdal and J. Surdej. *Rep Prog Phys* **56**, 117 (1994).

[187] R. Di Stefano and A.A. Esin, *Astrophys Journ* **448**, L1 (1995).

[188] R. D. Blandford and R. Narayan, *Astrophys Journ* **310**, 568 (1986).

[189] R. D. Blandford and R. Narayan, *Ann Rev Ast Ast* **30**, 311 (1992).

[190] J. A. Tyson, F. Valdes, and R. A. Wenk. *Astrophys Journ* **349** L1 (1990).

[191] C. C. Dyer and L. M. Oattes, *Ap. J.* **326**, 50 (1988).

[192] J. A. Tyson. In Gravitational lensing, ed. Y Mellier, B Fort and G Soucail. Springer Lecture Notes in Physics Volume 360 (Springer, 1990), 230.

[193] G. F. R. Ellis: *Ann New York Acad Sci* **336**, 130-160 (1980).

[194] M. S. Turner. *Physics World*, September 1966, p. 35.

[195] J. Wambsganss, H.J. Witt, and P. Schneider. MPA Preprint 623 (1991).

[196] W. L. Ames and K. S. Thorne *ApJ* **151**, 659 (1968).

[197] Y. Funato, J. Makino and T. Ebisuzaki. *Astrophys Journ* **424**, L17 (1994)

[198] Y. Funato, J. Makino & T. Ebisuzaki, Ap.J, **424**, L17 (1994).

[199] C. C. Dyer and L. M. Oattes, *Ap. J.* **326**, 50 (1988).

[200] S. J. Maddox *et al.*, M. N. R. A. S **249**, 1 (1991).

[201] J.P. Ostriker, Ann. Rev. Astro. & Astrophys. **31**, 689, (1993).

[202] M. S. Turner, *Phys Rev.* D **48**, 5539 (1993).

[203] R. Crittenden *et al.*, *Phys. Rev. Lett.*, **71**, 324, (1994).

[204] A. Babul and M. H. Lee, *M. N. R. A. S.* **250**,407 (1991).

[205] J. R. Bond *et al*, *Phy. Rev. Lett.*, **72**, 13 (1993).

[206] G. F. R. Ellis and W. R. Stoeger, *Classical Quant. Grav.* **4**, 1697 (1987).

[207] W. R. Stoeger, G. F. R. Ellis and C. Xu, Phys. Rev. D **49**, 1845 (1993).

[208] A. R. Liddle and D. Lyth, *Phys. Rep.* **231**, 1 & 2, 1 (1993).

[209] G. F. R. Ellis and T. Rothman, *American J. Phys.*, **61**, 93 (1993).

[210] M. Sasaki, *Prog. Theor. Phys.* **90**, 4, 753 (1993).

[211] J. A. Peacock in *New Insights into the Universe* Proceedings of the 1991 València Summer School, eds. V. J. Martínez, M. Portilla & D. Sàez, Springer Verlag

[212] G.F.R. Ellis, B.A.C.C. Bassett and P.K.S. Dunsby, *UCT preprint*, (1996) .

[213] N. Mustapha, B.A. Bassett, C.W. Hellaby and G.F.R. Ellis, UCT Preprint, *submitted to Class. Quant. Grav.* (1996)

[214] G.F.R. Ellis and D.M. Solomans, *UCT preprint*, (1996).

[215] C.C. Dyer and R.C. Roeder, *Astrophys. J.*, **174**, L115, (1972).

[216] S. Weinberg, *Astrophys. J.*, **208**, L1, (1976).

[217] R. Bar-Kana, astro-ph/9511056, Ap. J, Sept. 10 (1996)

[218] J. Wambsganss *et al*, *in press*, Ap.J L, astro-ph/9607084 (1996)

[219] R. Kantowski *et al*, Ap. J, **447**, 35 (1995)

[220] E.V. Linder, *Astron. and Astrophys.*, **206**, 190, (1988).

[221] G. Lemaître, *Ann. Soc. Scient. Bruxelles*, **A53**, 51, (1933).

[222] R.C. Tolman, *Proc. Nat. Acad. Sci.*, **20**, 169, (1934).

[223] H. Bondi, *Mon. Not. Roy. Astron. Soc.*, **107**, 410, (1947).

[224] R. Maartens, N.P. Humphreys, D.R. Matravers and W. Stoeger, *Class. Q. Grav.* **13**, 253-264, (1996). (gr-qc/9511045).

[225] N.P. Humphreys, R. Maartens and D.R. Matravers, *Portsmouth University preprint* RCG 96/1, (1996).

[226] G. Darmois, *Mémorial des Sciences Mathematiques*, Fasc.**25**, Gaulthier-Villars, Paris, (1927).

[227] W. Israel, *Il Nuovo Cim.*. **44B**, 1 (1966), and errata in *ibid* **48B**, 463, (1967).

[228] C. Hellaby and K. Lake, *Astrophys. J.*, **290**, 381-9, (1985), and errata in *ibid* **300**, 461, (1986).

[229] B. A. Bassett, *submitted to Letters in Applied Mathematics*, (1996)

[230] M. Rosenlicht, Am. Math. Monthly, Nov., 1972, 963

[231] M. Rosenlicht, Pacific Journal of Mathematics, **54**, 153-161 (1968); *ibid* 65, 485-492 (1976)

[232] I.S. Gradshteyn & I.M. Ryzhik, *Tables of Integrals, Series and Products*, Academic Press (1980)

[233] M. Abramowitz & I.A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, (1965)

[234] W.H. Press & P. Schecter, Formation of galaxies and clusters of galaxies by self-similar gravitational condensation, Ap.J, **187**, 425 (1974)

[235] J. Crank, *The Mathematics of Diffusion*, Oxford University Press, New York (1975),