Lecture Notes in Physics

109

Physics of the Expanding Universe

Cracow School on Cosmology
Jodłowy Dwór, September 1978
Poland

Edited by M. Demiański

Springer-Verlag
Berlin Heidelberg New York 1979

Editor

M. Demiański
Uniwersytet Warszawski
Instytut Fizyki Teoretycznej
ul. Hoża 69
00-681 Warszawa
Poland

ISBN 3-540-09562-4 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-09562-4 Springer-Verlag New York Heidelberg Berlin

Printed in Germany

Printing and binding: Beltz Offsetdruck, Hemsbach/Bergstr.
2153/3140-543210

FOREWARD

The Cracow School on Cosmology was held in Jodłowy Dwór and it was devoted to discussions of physical processes playing an important role in development and revealing the structure of universe. Methods used to study distribution of galaxies were also presented. I regret but I failed to persuaide Drs. M. Kalinkov and A. Kruszewski to write up their lectures.

The School was sponsored by Committee of Physics of the Polish Academy of Sciences, Jagiellonian University and Polish Astronomical Society. I would like to thank all those Institutions for their help.

I am deeply grateful to all lecturers and participants for creating a very good atmosphere and to Drs. P. Flin, Z. Klimek and L. Sokołowski for organizing the School.

I would like to dedicate the Proceedings to the memory of Dr. Z. Klimek who tragically died a few days before beginning of the School.

M. Demiański

TABLE OF CONTENTS

THE MATHEMATICS OF ANISOTROPIC SPATIALLY-HOMOGENEOUS COSMOLOGIES

Malcolm MacCallum
Department of Applied Mathematics, Queen Mary College, London

1. Introduction

1.1. Motivation

General-relativistic cosmology was for many years concerned almost entirely with the simplest possible models. These are the models which are both isotropic, i.e. in which all spatial directions are equivalent, and spatially-homogeneous, i.e. all points in space at a given time are equivalent. As we shall see later (Section 3.3), the condition of isotropy at every point leads uniquely to a certain metric form, a theorem due originally to Robertson and Walker. This form contains a function R(t) to be determined from Einstein's equations, and the first to do so, for dust, was Friedman. I shall therefore call the isotropic models Friedman-Robertson-Walker (FRW) models.

Until relatively recently, cosmological debate centred on comparing the FRW models with Robertson-Walker metrics of other gravitational theories. Strictly, this began with the comparison between Einstein's static universe, a solution of general relativity with a cosmological constant Λ, and the expanding solutions, the first of which was de Sitter's (also with cosmological constant) but later those of Friedman (with $\Lambda = 0$). Hubble's law, relating the redshifts and magnitudes of galaxies, first published in 1929, gave the advantage to expanding models, but unfortunately led, in the FRW models, to an age of the universe much shorter than the known age of the Earth. Various resolutions of this paradox, the most famous being "steady-state" theory, were attempted, as described, for example, by Bondi [93]. These controversies were resolved by

1) the revisions of the distance scale, leading to a much smaller Hubble constant and longer age for the universe,
2) the radio source counts, demonstrating that the universe really has evolved,
3) the discovery, in 1965, of the microwave background radiation, which gave powerful support to the hypothesis of an initial "big-bang" such as is found in FRW models.

The FRW models appear to fit the broad features of the present-day universe very well. Moreover, the advances in experimental testing of gravity in the last ten years, including the careful study of previously unconsidered effects by Thorne, Will, Ni, Nordvedt and others (see the reviews [94, 95] by Will), have led to a situation where general relativity seems confirmed as the most aesthetically pleasing theory fitting the known experimental facts. In this situation, the attention of cosmologists turned, in the 1960s to some new questions.

The point of these questions is that general relativity does not give a unique prediction for the universe. The governing equations must be supplemented by initial conditions, boundary conditions, symmetry conditions, and other restrictions in order to yield definite solutions, and there are an infinity of general-relativistic cosmologies. For various reasons, which I shall try to outline, attempts have been made to compare FRW models with these other, less symmetrical, models. The easiest models to consider are those which share with the FRW models the property of spatial-homogeneity. These are the models which form the subject of this course, and as those attending will discover there is a formidable body of literature about these models. It has been reviewed before, e.g. [96, 11, 97, 98], of course.

The philosophical reasons for considering non-FRW models were set out by Misner [99]. They are essentially that the Friedman models offer no explanation for the observed symmetry, and in particular that regions now observable (by the microwave radiation) could not have been in causal contact at time of emission, so that the symmetry really seems to be imposed, rather than natural. Misner suggested a programme of "chaotic cosmology", to test the hypothesis that arbitrary initial conditions would, by the operation of various physical processes, always reduce to the present-day observed universe.

A second aspect is that although FRW models start from a big-bang, thus satisfying the singularity theorems [45] which strongly indicate such an origin for our universe, they do not exhibit the most general types of singularity [100]. In particular, small perturbations of FRW models exist, i.e. small at some time t after the big-bang, such that they grow as the singularity is approached. The singularity structure is therefore unstable and the FRW initial conditions are far from general, being, in some ill defined sense, isolated in the space of solutions.

A further impetus to study of non-FRW models came from the fact that small perturbations caused by random statistical fluctuations in FRW models do not appear to grow fast enough for this to provide a sat-

isfactory account of galaxy formation.

It may seem strange, however, to pass from the FRW models to spatially-homogeneous anisotropic models. We can directly test, in a number of ways, the isotropy of the universe about us. Whether or not to abandon faith in isotropy is really an experimental question. However, it is almost impossible to test homogeneity, because we see distant regions as they were a long time ago, and in order to compare them with the present-day we must find the appropriate evolution to obtain the present-day parameters of those distant regions. This may well lead us into a circular argument.

Belief in homogeneity is really the outcome of a long series of reverses for a geocentric point of view. Briefly these were a) Copernicus' 1543 proposal that the Earth is not the center of the universe, b) Shapley's 1918 discovery that the Sun is not at the center of our Galaxy, c) Hubble's 1924 confirmation that the "island nebulae" were other galaxies and d) Baade's 1952 revision of the distance scale showing that ours is not the largest galaxy in the universe. The consequence is a widely-held belief, known as the "Copernican Principle", that the Earth is in no special place in the universe. Thus if we see isotropy, everybody must see isotropy, or if we see a linear Hubble law, so must everybody. Either of these situations leads to homogeneity in space.

The only attempts at direct testing of homogeneity use the distribution of galaxies, and since the galaxies appear to be clustered on scales which may be very large indeed, the outcome of these tests is disputed [101 - 104].

The principal advantage of the spatially-homogeneous models is that the physical variables depend only on time. Thus Einstein's equations, and the other governing equations, reduce to ordinary differential equations. My own view, however, is that the real resolution of difficulties of cosmology may lie with the consideration of inhomogeneous models.

Let me pass on to review briefly the question of the abandonment of isotropy. The early reasons for this, such as apparently low abundances of primordial helium in some stars, are no longer believed, but there are other indications which also must be considered. The various tests are

1. The distribution of galaxies. Some authors, especially de Vaucouleurs, believe this to be anisotropic.

2. The Hubble constant may contain an anisotropy associated with the Virgo supercluster, if that exists [105]. Rubin et al. [106] have

found anisotropies in the redshifts of ScI galaxies consistent with a velocity of 454 ± 125 km/sec in the direction $l = 163^{\circ}$, $b = -11^{\circ}$, but Schechter [107] has shown that the conclusion depends critically on the method of data analysis. From an independent sample he obtained a velocity 346 ± 76 km/sec in the direction of $l = 72^{\circ}$, $b = 28^{\circ}$. Another group finds anisotropies associated with the passage of light through clusters of galaxies [108].

3. Numerous studies have established that, contrary to earlier claims, there is no detectable positive anisotropy in the distribution or properties of radio sources (see e.g. [109]).

4. The cosmic X-ray background is isotropic to less than 5% [110].

5. The cosmic microwave background gives the most accurate tests of isotropy. Until last year no indication of anisotropy existed (see e.g. [111]) but now [112] Smoot et al. have found a velocity 603±60 km/sec towards $l = 261^{\circ}$, $b = 33^{\circ}$. It should be noted that this is not compatible with the present Hubble law measurements.

6. A cosmic magnetic field, which would break isotropy, has been proposed, but its existance is very doubtful. Tests are difficult because of the masking of effects by fields in our Galaxy and in the sources [113, 114]. If it exists, it could be related [115] to the various claims of anisotropy in orientation of galaxies and radio sources, [116 - 119] though these have been themselves firmly established [118] except perhaps within small regions [118 - 123].
There is thus no strong case for anisotropic models, but some of the data, if confirmed, could provide such evidence.

1.2. Aim of the following chapters

I was asked to talk about the classification and evolution of the models to be considered, but some of the other speakers will be concerned with certain of the more physical aspects. In particular Dr. Caderni will talk about viscous processes and Prof. Zeldovich about quantum effects. Dr. Partridge's lectures contain a survey of the recent data on the microwave background, while the galaxy distribution, mentioned above, is covered by Dr. Dautcourt.

I have therefore interpreted my task as that of providing the pure mathematical context within which the physics can be set. I do not aim to deal at all with the thermodynamic or kinetic theory aspects of the models, the calculation of the effects of dissipative processes, the introduction of quantum fields and particle creation, the conclusions to be drawn concerning isotropy from the element abundances and

microwave background, or any of the other exotic and exciting physics whose characteristics can be discussed in the models. (I may break these restrictions by giving a brief personal view at the end of my course.) Instead I shall focus on what I feel to be the equally exciting, if not exotic, mathematics required to set up the metrics, to elucidate the behaviour of the field equations and their principal characteristics, and to show how the geometry of the models can itself have dynamical importance, and impose restrictions on the matter dynamics. Chapter 2 of these notes is therefore devoted an introduction to symmetry groups of metrics and Chapter 3 to the construction of the metrics we wish to consider, without the use of Einstein's equations.

The Einstein equations are computed in Chapter 4, and their structure, as a system of ordinary differential equations, analyzed in terms of degrees of freedom, reducibility to various simplified forms, influence on matter content and known exact solutions. Chapter 5 deals with some qualitative effects of evolutions, particularly the dynamical effects of the geometry near to the singularity and in isotropization.

In these notes I have adopted the following conventions. Greek indices may run from 1...n for any n, but when used in a space-time run from 1 to 4 while Latin indices then run from 1 to 3. x^4 is the time coordinate. The signature is + 2 for Lorentz metrics and the Ricci identity reads

$$\nabla_{\underset{\sim}{X}_\alpha}\nabla_{\underset{\sim}{X}_\beta}\underset{\sim}{X}_\gamma - \nabla_{\underset{\sim}{X}_\beta}\nabla_{\underset{\sim}{X}_\alpha}\underset{\sim}{X}_\gamma - \nabla_{[\underset{\sim}{X}_\alpha,\underset{\sim}{X}_\beta]}\underset{\sim}{X}_\gamma = R^{\delta}{}_{\gamma\alpha\beta}\ \underset{\sim}{X}_\delta$$

for any vector basis $\{\underset{\sim}{X}_\alpha\}$, where $\nabla_{\underset{\sim}{Y}}\underset{\sim}{Z}$ denotes the covariant derivative of $\underset{\sim}{Z}$ in the $\underset{\sim}{Y}$ direction. The Ricci tensor and Ricci scalar are

$$R_{\alpha\beta} = R^{\delta}{}_{\alpha\delta\beta} \quad , \quad R = R^{\alpha}{}_{\alpha}$$

and the units are chosen so that Einstein's equations read

$$G_{\alpha\beta} + \Lambda g_{\alpha\beta} = T_{\alpha\beta}$$

where $G_{\alpha\beta} = R_{\alpha\beta} - \frac{1}{2} R g_{\alpha\beta}$ is the Einstein tensor.

A comma between indices denotes partial differentiation, with respect to subsequent indices, a semi-colon covariant differentiation. Indices in square brackets are to be skewed over, those in round brackets to be symmetrized.

2. Introduction to transformation groups and isometries

2.1. The Lie derivative

Let $\underset{\sim}{Y}$ be a vector field on a manifold M, and suppose its integral curves are $\gamma(u)$, u being the coordinate such that $\underset{\sim}{Y} = \partial/\partial u$. We can then consider the transformations Φ_u: $\gamma(u) \longmapsto (u + u)$ where u has a fixed value for each Φ_u. These transformations simply carry each point of M a parameter distance u along the integral curve passing through that point.

Since Φ_u maps M onto itself, it maps geometric object fields on M to similar geometric object fields. (A geometric object is an entity which has well-defined components in any coordinate system, and a well-defined law of transformation for these components under any coordinate change. Tensors are geometric objects, but so too are such quantities as the connection, which do not obey the tensor transformation law. The action of Φ_u is called "Lie transport" or "dragging along".) One can now compare the value of $\Phi_u \underset{\sim}{G}$ at a point p with the value of $\underset{\sim}{G}$ at p, for any geometric object $\underset{\sim}{G}$. (Here I use Φ_u rather inexactly to denote the map of geometric objects associated with Φ_u itself.)

The Lie derivative $\mathcal{L}_{\underset{\sim}{Y}} \underset{\sim}{G}$ of a geometric object $\underset{\sim}{G}$ with respect to $\underset{\sim}{Y}$ is defined to be

$$\mathcal{L}_{\underset{\sim}{Y}} \underset{\sim}{G} = \lim_{u \to 0} \left(\frac{\underset{\sim}{G} - \Phi_u \underset{\sim}{G}}{u} \right) .$$

One can evaluate the components of $\mathcal{L}_{\underset{\sim}{Y}} \underset{\sim}{G}$ quite easily, at a point p, by taking a small u and using a coordinate system $\{x^\mu\}$. Let the point p have coordinates p^μ and the vector field $\underset{\sim}{Y}$ have components Y^μ. The point q such that $\Phi_u q = p$ has coordinates $p^\mu - uY^\mu(p)$, up to order u. Using this formula for a neighbourhood of p we define new coordinates $x'^\mu = x^\mu - uY^\mu(x)$. $\Phi_u \underset{\sim}{G}$ has the same components at x'^μ (in the $\{x'^\mu\}$ system) as it had at x'^μ in the $\{x^\mu\}$ system. The components of $\underset{\sim}{G}$ in the x'^μ system can be evaluated and compared with the components of $\Phi_u \underset{\sim}{G}$, and so on. Let us carry this through for a vector field $\underset{\sim}{Z}$.

The components of $\underset{\sim}{Z}$ at q in the $\{x^\mu\}$ system (i.e. of $\Phi_u \underset{\sim}{Z}$ at p in the $\{x'^\mu\}$ system) are $Z^\mu(p) - uZ^\mu{}_{,\nu}\, Y^\nu(p)$, to order u. The components of $\underset{\sim}{Z}$ at p in the $\{x'^\mu\}$ system are $Z^\nu \frac{\partial x'^\mu}{\partial x^\nu} = Z^\nu(\delta^\mu_\nu - uY^\mu{}_{,\nu})$ to order u. Thus

$$(\mathcal{L}_{\underset{\sim}{Y}} \underset{\sim}{Z})^\mu = Z^\mu{}_{,\nu} Y^\nu - Z^\nu Y^\mu{}_{,\nu} \quad . \qquad (2.1a)$$

This has the same components as $[\underset{\sim}{X}, \underset{\sim}{Z}]$, the commutator of the two vector fields. (The commutator can be understood by treating $\underset{\sim}{X}$ as a differential operator $X^\mu \partial/\partial x^\mu$ and defining $[\underset{\sim}{X},\underset{\sim}{Z}]$ to be the operator such that for any scalar function ψ , $[\underset{\sim}{X}, \underset{\sim}{Z}]\psi = \underset{\sim}{X}(\underset{\sim}{Z}\psi)-\underset{\sim}{Z}(\underset{\sim}{X}\psi)$.)

Thus

$$\mathcal{L}_{\underset{\sim}{X}}\underset{\sim}{Z} = [\underset{\sim}{X}, \underset{\sim}{Z}] . \qquad (2.1b)$$

Since the Lie derivative can easily be shown to obey the normal rules of differentiation (e.g. Leibnitz' rule), and it is also easy to see that for a scalar

$$\mathcal{L}_{\underset{\sim}{X}} \psi = \underset{\sim}{X}\psi ,$$

one can readily deduce from (2.1) the effect of $\mathcal{L}_{\underset{\sim}{X}}$ on any tensor. For example, for a differential form, with components ω_μ , we know that $\omega_\mu Z^\mu$ is a scalar, if $\underset{\sim}{Z}$ is an arbitrary vector field. Thus

$$\begin{aligned}\mathcal{L}_{\underset{\sim}{X}}(\omega_\mu Z^\mu) &= \omega_{\mu,\nu} Z^\mu X^\nu + \omega_\mu Z^\mu{}_{,\nu} X^\nu \\ &= (\mathcal{L}_{\underset{\sim}{X}} \underset{\sim}{\omega})_\mu Z^\mu + (\mathcal{L}_{\underset{\sim}{X}}\underset{\sim}{Z})^\mu \omega_\mu\end{aligned}$$

so

$$(\mathcal{L}_{\underset{\sim}{X}} \underset{\sim}{\omega})_\mu Z^\mu = (\omega_{\mu,\nu}X^\nu + X^\nu{}_{,\mu}\omega_\nu) Z^\mu$$

using (2.1). Since $\underset{\sim}{Z}$ is arbitrary this implies

$$(\mathcal{L}_{\underset{\sim}{X}} \underset{\sim}{\omega})_\mu = \omega_{\mu,\nu} X^\nu + X^\nu{}_{,\mu}\omega_\nu$$

For a tensor with two covariant indices (i.e. components $a_{\mu\nu}$) we can show similarly that

$$(\mathcal{L}_{\underset{\sim}{X}}\underset{\sim}{a})_{\mu\nu} = a_{\mu\nu,\rho} X^\rho + a_{\mu\rho} X^\rho{}_{,\nu} + a_{\rho\nu} X^\rho{}_{,\mu} \qquad (2.2)$$

Finally note that when we are dealing with a Riemannian manifold, with metric $\underset{\sim}{g}$ say,

$$\begin{aligned}(\mathcal{L}_{\underset{\sim}{X}}\underset{\sim}{Z})^\mu &= Z^\mu{}_{;\nu} X^\nu - X^\mu{}_{;\nu} Z^\nu \\ (\mathcal{L}_{\underset{\sim}{X}} \underset{\sim}{\omega})_\mu &= \omega_{\mu;\nu} X^\nu + X^\nu{}_{;\mu}\omega_\nu \qquad (2.3)\end{aligned}$$

and

$$(\mathcal{L}_{\underset{\sim}{X}}\underset{\sim}{g}) = X_{\mu;\nu} + X_{\nu;\mu} .$$

These results follow from the symmetry of the connection and the property $g_{\mu\nu;\rho} = 0$.

2.2. Lie groups and Lie algebras

The topic of this section is rich enough to fill at least one complete book, more likely several, and I cannot possibly hope to give here more than the sketchiest of outlines. Fortunately there are many available texts for the interested reader to consult; I myself learnt the subject from the book of Cohn [1].

A Lie group G is a group (in the usual sense of algebra) and also a differential manifold such that the map $G \times G \longmapsto G$ given by the algebraic product $(a,b) \longmapsto ab$ is differentiable. (Strictly, it must be twice differentiable, and it must then be analytic. See Cohn [1] pages 44-47. In the usual tradition of British applied mathematics I shall ignore all such analytical problems on the grounds that we can always approximate the physical situation by a model as smooth as we like.)

If $a \in G$, we define the right translation R_a and left translation L_a associated with a to be the maps of G into itself defined by

$$bL_a = ab, \qquad bR_a = ba.$$

(For clarity in this section we follow the algebraist's convention of writing the map on the right, so that the composition of two maps f and g, where g is performed after f, will be written fg automatically.) Since $aR_bR_c = abc = aR_{bc}$ we see that the set of right translations forms a group isomorphic to G itself. The left translations give $L_bL_c = L_{cb}$, so they are in fact a group algebraically dual to G: as we shall see, this group is also isomorphic to G. The left translations commute with right translations, i.e. $aL_bR_c = aR_cL_b$.

A left-invariant vector field on G is defined to be one which is invariant under left translation, i.e. if $\underset{\sim}{y}$ is the vector field, its value at ab is given by

$$\underset{\sim}{y}(ab) = \underset{\sim}{y}(b)(L_a)_*$$

where $(L_a)_*$ is the map of vectors associated with L_a. Given the value $\underset{\sim}{y}(e)$ at the identity e of G, this defines $\underset{\sim}{y}$ uniquely at all points (by $\underset{\sim}{y}(a) = \underset{\sim}{y}(e)(L_a)_*$, which is easily shown to be left-invariant), and conversely $\underset{\sim}{y}$ uniquely defines $\underset{\sim}{y}(e)$. This shows that G has the same dimension at all points (since $(L_{a^{-1}})_*$ and $(L_a)_*$ give an isomorphism between $\tau_a(G)$ and $\tau_e(G)$ where $\tau_p(M)$ denotes the space of vectors tangent to M at p).

From a left-invariant vector field, we can construct a mapping

Φ_u: G$\mapsto$G as in Section 2.1. It is clear that this map commutes with left translations, by its construction. Then if $e\Phi_u = b$, we find that

$$a\Phi_u = eL_a\Phi_u = e\Phi_u L_a = bL_a = ab.$$

So $\Phi_u = R_b$. Thus we deduce that the left-invariant vector fields represent infinitesimal right translations. Similarly, right invariant vector fields represent infinitesimal left translations.

The left-invariant vector fields form a vector space of the same dimension as G (since the correspondence of $\underset{\sim}{y}$ with $\underset{\sim}{y}(e)$ is a linear map). The commutator of two left-invariant vector fields is left-invariant, since for any map f: M$\mapsto$N where M, N are differential manifolds, the corresponding map f_* of vectors gives

$$[\underset{\sim}{X}, \underset{\sim}{Y}]f_* = [\underset{\sim}{X}f_*, \underset{\sim}{Y}f_*] .$$

This last statement follows from taking an arbitrary scalar function ψ on N, and using the definition $fo(\psi(\underset{\sim}{X}f_*)) = (fo\psi)\underset{\sim}{X}$ then

$$\begin{aligned} fo(\psi[\underset{\sim}{X},\underset{\sim}{Y}]f_*) &= (fo\psi)[\underset{\sim}{X},\underset{\sim}{Y}] = ((fo\psi)\underset{\sim}{Y})\underset{\sim}{X} - ((fo\psi)\underset{\sim}{X})\underset{\sim}{Y} \\ &= (fo(\psi(\underset{\sim}{Y}f_*))\underset{\sim}{X} - (fo(\psi(\underset{\sim}{X}f_*))\underset{\sim}{Y}) \\ &= fo((\psi(\underset{\sim}{Y}f_*))\,\underset{\sim}{X}f_*) - fo((\psi(\underset{\sim}{X}f_*))\underset{\sim}{Y}f_*) \\ &= fo(\psi[\underset{\sim}{X}f_*, \underset{\sim}{Y}f_*]) \end{aligned}$$

The commutator $[\underset{\sim}{X},\underset{\sim}{Y}]$ has the properties

$$[\underset{\sim}{X},\underset{\sim}{X}] = 0$$

$$[\underset{\sim}{X},\underset{\sim}{Y}] = -[\underset{\sim}{Y},\underset{\sim}{X}]$$

$$[\underset{\sim}{X},[\underset{\sim}{Y},\underset{\sim}{Z}]] + [\underset{\sim}{Y},[\underset{\sim}{Z},\underset{\sim}{X}]] + [\underset{\sim}{Z},[\underset{\sim}{X},\underset{\sim}{Y}]] = 0 \qquad (2.4)$$

The equation (2.4) is known as the Jacobi identity. Any vector space with a product linear in $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ and obeying the last three equations forms, by definition, a Lie algebra. Thus we see that a Lie group has associated with it a unique Lie algebra. It is possible to show that every Lie algebra defines a unique Lie group, G′ say, such that each other Lie group with the same Lie algebra is an image of G′ under some homomorphism; this theorem is far too difficult for these lectures. The different groups with the same Lie algebra differ only in their (global) topological properties, e.g. connectivity. Thus the algebraic structure of G is defined (up to these topological considerations) entirely by that of its Lie algebra. It is usual to express the latter by taking a basis $\{\underset{\sim}{X}_\alpha\}$ of the Lie algebra, where $\alpha = 1,$

2,3...r and r is the dimension of G. Then $[\underset{\sim}{X}_\alpha, \underset{\sim}{X}_\beta]$ is in the Lie algebra and so

$$[\underset{\sim}{X}_\alpha, \underset{\sim}{X}_\beta] = C^\gamma{}_{\alpha\beta}\, \underset{\sim}{X}_\gamma \tag{2.5}$$

for some constants $C^\gamma{}_{\alpha\beta}$. The $C^\gamma{}_{\alpha\beta}$ are called the structure constants of the Lie algebra, or equivalently, of its associated Lie group(s). The Jacobi identity (2.4) implies

$$C^\alpha{}_{\beta[\gamma} C^\beta{}_{\delta\varepsilon]} = 0, \tag{2.6}$$

and clearly $C^\alpha{}_{\beta\gamma} = - C^\alpha{}_{\gamma\beta}$.

Now let us return to the right-invariant vector fields. They also form a Lie-algebra of dimension r, as is readily shown by interchanging "right" and "left" in the above arguments, and we shall write a basis for it as $\{\underset{\sim}{W}_\alpha\}$. The fact that left and right translations commute leads to

$$[\underset{\sim}{X}_\alpha , \underset{\sim}{W}_\beta] = 0. \tag{2.7}$$

We have also

$$[\underset{\sim}{W}_\alpha , \underset{\sim}{W}_\beta] = D^\gamma{}_{\alpha\beta}\, \underset{\sim}{W}. \tag{2.8}$$

Now since, at any point a of G, $\underset{\sim}{X}_\alpha(a)$ and $\underset{\sim}{W}_\beta(a)$ are bases of $\tau_a(G)$, we must have

$$\underset{\sim}{W}_\alpha = M_\alpha{}^\gamma \underset{\sim}{X}_\gamma$$

for some non-singular position-dependent matrix $M_\alpha{}^\gamma$ (2.7) implies

$$M_\alpha{}^\gamma C_{\gamma\beta}{}^\delta - \underset{\sim}{X}_\beta(M_\alpha{}^\delta) = 0$$

and then (2.8) yields

$$\begin{aligned}
[\underset{\sim}{W}_\alpha , \underset{\sim}{W}_\beta] &= [M_\alpha{}^\gamma \underset{\sim}{X}_\gamma , M_\beta{}^\delta \underset{\sim}{X}_\delta] \\
&= (M_\alpha{}^\gamma M_\beta{}^\delta [\underset{\sim}{X}_\gamma , \underset{\sim}{X}_\delta] + (M_\alpha{}^\gamma \underset{\sim}{X}_\gamma (M_\beta{}^\delta)) \underset{\sim}{X}_\delta \\
&\quad - M_\beta{}^\delta \underset{\sim}{X}_\delta (M_\alpha{}^\gamma)) \underset{\sim}{X}_\gamma \\
&= (M_\alpha{}^\gamma M_\beta{}^\delta C^\varepsilon{}_{\gamma\delta} + M_\alpha{}^\gamma M_\beta{}^\delta C^\varepsilon{}_{\delta\gamma} - M_\beta{}^\delta M_\alpha{}^\gamma C^\varepsilon{}_{\gamma\delta}) \underset{\sim}{X}_\varepsilon \\
&= - (M_\alpha{}^\iota M_\beta{}^\delta C^\varepsilon{}_{\gamma\delta} M^{-1}{}_\varepsilon{}^\varphi) \underset{\sim}{W}_\varphi
\end{aligned}$$

where $M^{-1}{}_\varepsilon{}^\varphi$ is the inverse matrix of $M_\alpha{}^\beta$. By taking the bases to be such that $M_\alpha{}^\gamma = - \delta_\alpha^\gamma$ at e, we find

$$D^\alpha{}_{\beta\gamma} = C^\alpha{}_{\beta\gamma} \tag{2.9}$$

Thus the Lie algebras are isomorphic, and so are the Lie groups. An

isomorphism is given by $f: R_a \longmapsto L_{a^{-1}}$, for then $R_{ab} \longmapsto L_{(ab)^{-1}} = L_{b^{-1}a^{-1}} = L_{a^{-1}}L_{b^{-1}} = f(R_a)f(R_b)$. It is more usual to take $M_\alpha{}^\gamma = \delta_\alpha^\gamma$ at e, with the results that

$$D^\alpha{}_{\beta\gamma} = - C^\alpha{}_{\beta\gamma} \tag{2.10}$$

One can define subgroups of G, and subalgebras of its Lie algebra $\hat{G}$, in the usual way. It is then possible to prove that if $\hat{H}$ is any subalgebra of $\hat{G}$, it corresponds to a subgroup H of G which is an analytic sub-Lie-group (e.g. Cohn [1], page 120). If H is a normal subgroup and closed (in the sense of point-set topology) in G, then one can form the quotient group G/H (in the usual way, i.e. from cosets of H in G) and its Lie algebra is isomorphic to $\hat{G}/\hat{H}$ (Cohn [1], page 132). This implies that $\hat{H}$ is an ideal of $\hat{G}$, i.e. that $\hat{H}$ is a linear subspace of the vector space $\hat{G}$ and also that if $\underset{\sim}{X} \in \hat{G}$ and $\underset{\sim}{Z} \in \hat{H}$ then $[\underset{\sim}{X}, \underset{\sim}{Z}] \in \hat{H}$. Finally one can show that if $\hat{H}$ is an ideal of $\hat{G}$, then its corresponding subgroup H is normal in G (Cohn [1], page 138).

2.3. Bianchi classification of real Lie algebras

It is possible to systematically list all real Lie algebras which are non-isomorphic. For the 3-dimensional algebras this was first done by Bianchi [2], and it is this classification that is relevant to spatially-homogeneous cosmologies. Let me first, however, give the result for two-dimensional groups. The classification examines the commutators. These themselves form a Lie algebra, which is a subalgebra of $\hat{G}$ called the (first) derived algebra of $\hat{G}$. In the case of a two-dimensional group there is only one commutator to consider. If this is zero, the group is abelian and is called type G_2I; any basis obeys

$$[\underset{\sim}{X}_1, \underset{\sim}{X}_2] = 0 \tag{2.11}$$

If the commutator is not zero, the derived algebra is one-dimensional and we can choose $\underset{\sim}{X}_1$ to lie in it. Then $|\underset{\sim}{X}_1, \underset{\sim}{X}_2| = \alpha \underset{\sim}{X}_1$ for some constant α, and by scaling $\underset{\sim}{X}_2$ we can arrange that

$$[\underset{\sim}{X}_1, \underset{\sim}{X}_2] = \underset{\sim}{X}_1. \tag{2.12}$$

This non-abelian group is classified as G_2II.

As with the G_2, Bianchi's method with the three-dimensional algebras was to consider first the dimension of the derived algebra and then to enumerate all possibilities. This gave him nine inequivalent types, of which type I is abelian, and has zero-dimensional Lie alge-

bra, types II and III have a one-dimensional derived algebra, types IV, V, VI and VII two-dimensional, and types VIII and IX three-dimensional. Types VI and VII in fact are one-parameter families of algebras, where certain values of the parameter are excluded because they yield types III and V instead. Bianchi's classification has been modified in recent years [3,4] and the present method is as follows.

Take any (positive definite) scalar product on $\hat{G}$, and suppose its components in the basis $\{X_\alpha\}$ are $g_{\alpha\beta}$. Let $\epsilon^{\alpha\beta\gamma}$ be the corresponding completely-skew pseudo-tensor. Then write

$$\frac{1}{2} C^{\alpha}{}_{\beta\gamma} \epsilon^{\beta\gamma\delta} = N^{(\alpha\delta)} + \epsilon^{\alpha\delta\eta} A_\eta \quad . \tag{2.13}$$

This defines the vector A_η (on $\hat{G}$) uniquely, since

$$A_\varsigma = \frac{1}{4} \epsilon_{\varsigma\alpha\delta} C^{\alpha}{}_{\beta\gamma} \epsilon^{\beta\gamma\delta} = \frac{1}{2} C^{\alpha}{}_{\varsigma\alpha}$$

and it defines $N^{\alpha\delta}$, which is symmetric, up to an overall (positive or negative) scale factor, since any two completely skew r-tensors on a vector space of dimension r are proportional. The Jacobi identity (2.6) is equivalent to

$$N^{\alpha\beta} A_\beta = 0. \tag{2.14}$$

The classification now gives two broad classes, Class A where $A_\beta = 0$, and Class B ($A_\beta \neq 0$), each divided into several types according to the rank and (the modulus of the) signature of $N^{\alpha\beta}$. Clearly in Class B the rank of $N^{\alpha\beta}$ is less than, or equal to, 2. When $A_\beta \neq 0$ there is a further invariant h, which can be defined by

$$(1 + h) C^{\alpha}{}_{\beta\alpha} C^{\delta}{}_{\gamma\delta} = -2h\, C^{\alpha}{}_{\delta\beta} C^{\delta}{}_{\alpha\gamma}.$$

This is the one parameter required to subclassify types VI and VII: when $h = 0$ and $A_\beta \neq 0$ we have type V. If $A_\beta = 0$ then $h = 0$ and we have types VI_0 and VII_0 which belong to Class A. When $h = -1$ we have type III. In general $h > 0$ gives type VII and $h < 0$ gives type VI.

By rotating the basis $\{X_\alpha\}$ (rotation being defined relative to the metric $g_{\alpha\beta}$ on $\hat{G}$) we can diagonalize the matrix $n^{\alpha\beta}$ so that $A_\beta = (A,0,0)$ and $N^{\alpha\beta} = \mathrm{diag}\,(N_1,N_2,N_3)$ and then by scaling the basis we can set the non-zero entries in $N_{\alpha\beta}$ to +1 or −1 as appropriate. In types IV and V we can also scale to set $A = 1$. In general $h = A^2/N_2N_3$ so the scaling gives $A = \sqrt{|h|}$. The resulting classification and canonical forms are shown in Table I.

The canonical form for the structure constants does not fix the basis uniquely. Suppose we consider an arbitrary change of basis, i.e.

take any linear transformation in the group GL(3) of transformations of a three-dimensional vector space. GL(3) is a nine-dimensional group, since it can be represented by (non-singular) arbitrary 3×3 matrices. The nine structure constants must satisfy three Jacobi identities (2.14). There are thus six arbitrary entries. In Class A these are the entries in the 3×3 symmetric matrix $N_{\alpha\beta}$, while in Class B they are the 3 entries in the vector A_β and the 3 entries in a symmetric 2×2 matrix transforming a plane complementary to A_β.

Table I. The Bianchi Classification

Class	A						B				
Type	I	II	VI_0	VII_0	VIII	IX	V	IV	III	VI_h	VII_h
Rank $N_{\alpha\beta}$	0	1	2	2	3	3	0	1	2	2	2
|signature $(N_{\alpha\beta})$|	0	1	0	2	1	3	0	1	0	0	2
A	0	0	0	0	0	0	1	1	1	-h	h
N_1	0	1	0	0	-1	1	0	0	0	0	0
N_2	0	0	-1	1	1	1	0	0	-1	-1	1
N_3	0	0	1	1	1	1	0	1	1	1	1

In types VIII, IX, VI_h and VII_h there are no further restrictions, so there are 6-dimensional sets of structure constants for each type here (5-dimensional in types VI_h and VII_h if the value of h is fixed). In types VI_0 and VII_0 the matrix $N^{\alpha\beta}$ must be singular, so there are 5 independent values of structure constants, and similarly in type IV the (2×2) matrix must be singular, leaving 5 freely specifiable structure constants. In types II and V only a vector (3 free constants) can be given, this being A_β in type V, and the first row (say) of $N^{\alpha\beta}$ in type II. In type I there are clearly no free constants. This information is summarized in Table II, where d is the number of freely specifiable constants giving a particular group type.

Table II. Freedom to specify structure constants for a given group type

Type	I	II	VI_0	VII_0	VIII	IX	V	IV	VI_h	VII_h
d	0	3	5	5	6	6	3	5	5(6)	5(6)

In Table II the 5(6) under types VI_h and VII_h reflects the ambiguity about whether h is to be specified or not. Using the terminology and results explained in the next section, it emerges that the stability group of the canonical form of the structure constants (within GL(3)) has dimension 9-d from equation (2.15). These considerations are implicit in some work by myself [5], stated by Collins and Hawking [6], and further explored by Siklos [7].

It is possible to classify higher-dimensional algebras but these results will not be required for our purposes in these lectures.

2.4. Groups of Transformations

Suppose we have a manifold M and a Lie group G_r of transformations of M onto itself. We write the abstract elements of G as a,b,... and the corresponding maps of M as $T_a, T_b, \ldots$, the action of T_a on M being given by $p \longmapsto pT_a$ where p is any point in M. (Strictly, the group of the T_a is homomorphic to, not identical with, the abstract group.) We must have

$$pT_aT_b = pT_{ab}.$$

If id_M is the transformation $p \longmapsto p$ for every $p \in M$ the group G is said to act <u>effectively</u> if and only if $T_a = id_M$ implies $a = e$, where e is the identity of G. If G does not act effectively, the subgroup N such that $T_a = id_M$ is clearly normal and so can be eliminated by passing to the transformation group G/N. From now on we shall be speaking only of effective transformation groups.

A given left-invariant vector field $\underset{\sim}{v}$ on G gives rise to a one-parameter subgroup of right translations $R_{b(u)}$ on G where $b(u) = e\Phi_u$. This gives rise to a one-parameter subgroup $T_{b(u)}$ of transformations of M and thus to a vector field $\underset{\sim}{V}$ on M. The vector fields defined in this way form a Lie algebra isomorphic to that of G itself. Since the construction gives the zero vector field on M only if $\underset{\sim}{v}$ is zero on G (because G is effective), we need only find some map $f : G \longmapsto M$ such that $\underset{\sim}{v}f_* = \underset{\sim}{V}$; the result then follows because f_* is linear and $[\underset{\sim}{X}f_*, \underset{\sim}{Y}f_*] = [\underset{\sim}{X},\underset{\sim}{Y}]f_*$. Actually we can do this only on submanifolds of M called the orbits of G. These are defined as follows: if $p \in M$, the <u>orbit</u> of G through p is the set $\{pT_a\}$ of all points pT_a, $a \in G$.

In an orbit we define $pT : G \longmapsto M$; $a \longmapsto pT_a$, taking a fixed p. The required map is then $(pT)_*$. To show this we first note that if the base point is q, we obtain

$$qT_bT_a = qT_{ba} = (ba)(pT) = (aL_b)(qT) = a(L_b(qT)).$$

Thus if we had chosen a different base point q, such that $p = qT_b$, then $(pT) = L_b(qT)$ and so $(pT)_* = (L_b)_*(qT)_*$. If $\underset{\sim}{y}$ is a left-invariant vector field we then have $\underset{\sim}{y}(pT)_* = \underset{\sim}{y}(L_b)_*(qT)_* = \underset{\sim}{y}(qT)_*$. Therefore we need only evaluate the effect of pT_* at e. If we consider the map $pT_b : R \longmapsto G \longmapsto M$ given by $u \longmapsto b(u) \longmapsto pT_{b(u)}$, we find $pT_{b(u)} = e\Phi_u(pT)$ and thus at p

$$\underset{\sim}{Y} = \underset{\sim}{1}(pT_b)_* = \underset{\sim}{1}(e\Phi_u)_*(pT)_* = \underset{\sim}{y}(pT)_* ,$$

using the chain rule.

Thus we have proved that on each orbit the vector fields $\underset{\sim}{Y}$ form a Lie algebra isomorphic to that of G, and this must be true at every point $p \in M$, since $p \in \{pT_a\}$.

We define the stability group of p to be the group $\{ a: a \in G$ and $pT_a = p\}$. If a is in the stability group of p and b is not then $(pT_b)T_{b^{-1}}T_aT_b = pT_aT_b = pT_b$ so $b^{-1}ab$ is in the stability group of pT_b. Thus the stability groups of different points in the same orbit are conjugate subgroups of G. The Lie algebra of the stability subgroup consists of those left-invariant vector fields such that $\underset{\sim}{y}(pT)_* = 0$ at p. Thus if r is the dimension of G, and s the dimension of the stability subgroup of p, then

$$r = m + s \qquad (2.15)$$

where m is the dimension of the space spanned by all $\underset{\sim}{y}(pT)_*$ at p. This is in fact the dimension of the orbit through p, since pT maps G onto the orbit. (Perhaps I should remark that although pT is a perfectly good map of manifolds it does not preserve the group structure of G, in general.)

A transformation group is said to be <u>transitive</u> on its orbits, <u>simply-transitive</u> if $s = 0$ and <u>multiply-transitive</u> if $s > 0$. It is possible for s to be different on different orbits, as in the case of rotation of a plane, where $s = 1$ at the centre of rotation and 0 elsewhere. However s is constant within an orbit because conjugacy is an isomorphism and the stability groups of different points are conjugate; hence we also find m, the dimension, is constant throughout an orbit.

In the case of a simply-transitive group, one can, by choosing a fixed p, make each orbit isomorphic to G itself using pT (assuming the global topologies are identical). In this case the product $(pT_a)(pT_b)$ can be unambiguously defined as pT_{ab} and pT now preserves the group structure. The map gives a map of the right-invariant vector fields $\underset{\sim}{w}$ to vector fields on the orbit. A basis $\{\underset{\sim}{W}_\alpha\}$ of right-invariant fields then gives a basis, which, with some loss of clarity,

we also denote $\{\underset{\sim}{W}_\alpha\}$, of vector fields on the orbit (a basis in the sense that any vector field can be written as $\sum_\alpha z^\alpha \underset{\sim}{W}_\alpha$ with non-constant z^α). By a similar abuse of language we shall use $\{\underset{\sim}{X}_\alpha\}$ for a basis both in G and in the orbit, using left-invariant fields on G. If we take any left invariant vector field $\underset{\sim}{v}$ with corresponding field $\underset{\sim}{V}$ on the orbit then $[\underset{\sim}{W}_\alpha, \underset{\sim}{V}] = 0$.

One can find canonical coordinates on G, for example, by finding, for a given point $a \in G$, the vector $\underset{\sim}{v} = \sum_\alpha x^\alpha \underset{\sim}{X}_\alpha$ such that $a = e\Phi_1$, Φ_u being defined by $\underset{\sim}{v}$, where $\underset{\sim}{X}_\alpha$ are some basis of left-invariant vector fields and x^α are taken as the coordinates of a. These same coordinates can be used on the orbits of simply-transitive groups, using an identification pT. The basis $\underset{\sim}{X}_\alpha$ can be chosen to give the canonical form of the structure constants. The origin of the coordinates will be at e in G or p in an orbit, and the basis vectors $\{\underset{\sim}{X}_\alpha\}$ will have the values $\frac{\partial}{\partial x^\alpha}$ at the origin. It is possible to choose a basis $\{\underset{\sim}{W}_\alpha\}$ of the right invariant fields in accordance with (2.10). Let us denote the duals of the $\underset{\sim}{W}_\alpha$ by ω^α. Then since $\mathcal{L}_{\underset{\sim}{X}_\beta} \underset{\sim}{W}_\alpha = [\underset{\sim}{X}_\beta, \underset{\sim}{W}_\alpha] = 0$, $\mathcal{L}_{\underset{\sim}{X}_\beta} \omega^\alpha = 0$ also. In table III I give values for $\{\underset{\sim}{X}_\alpha\}$, $\{\underset{\sim}{W}_\beta\}$ and ω^β for the Bianchi types with structure constants as listed in Table I. These forms depend first on the chosen canonical form of structure constants. They then depend on the choice among the (9-d)-dimensional set of bases $\{\underset{\sim}{X}_\alpha\}$ with the same structure constants (as described in section 2.3). There are in fact alternative ways of defining canonical x^α from the $\underset{\sim}{X}_\alpha$ (see e.g. Cohn [1], page 110). There is then the freedom to choose the $\{\underset{\sim}{W}_\alpha\}$ at e and the point p in the orbit. For these reasons, many forms can be found in the literature. With our conventions the freedom is restricted to the (9-d)-dimensional choice of basis $\{\underset{\sim}{X}_\alpha\}$ and, in the orbits of simply-transitive groups, the choice of p.

The resulting vector fields $\{\underset{\sim}{W}_\alpha\}$ generate another group of transformations on the orbits of a simply-transitive group, which is called the reciprocal group. It is the image of the group of left translations and is algebraically dual to the transformation group itself.

Having completed my remarks about purely algebraic aspects of the problem, I shall now revert, where appropriate, to writing maps on the left in accordance with the usual analyst's convention.

The most thorough treatment of groups of transformations is in Eisenhart's book [8]; I know of no comparable work using modern differential geometric notation.

Table III. Values for coordinate expressions for basis vectors and forms in each Bianchi type. Here $\partial_\alpha \equiv \partial/\partial x^\alpha$.

Type	I	II	IV	V
$\underset{\sim}{X}_\alpha$	∂_1	∂_1	$\partial_1 - x^2\partial_2 - (x^2+x^3)\partial_3$	$\partial_1 - x^2\partial_2 - x^3\partial_3$
	∂_2	∂_2	∂_2	∂_2
	∂_3	$\partial_3 + x^2\partial_1$	∂_3	∂_3
$\underset{\sim}{W}_\alpha$	∂_1	∂_1	∂_1	∂_1
	∂_2	$\partial_2 + x^3\partial_1$	$e^{-x^1}(\partial_2 - x^1\partial_3)$	$e^{-x^1}\partial_2$
	∂_3	∂_3	$e^{-x^1}\partial_3$	$e^{-x^1}\partial_3$
$\underset{\sim}{\omega}^\alpha$	dx^1	$dx^1 - x^3dx^2$	dx^1	dx^1
	dx^2	dx^2	$e^{x^1}dx^2$	$e^{x^1}dx^2$
	dx^3	dx^3	$e^{x^1}(x^1dx^2 + dx^3)$	$e^{x^1}dx^3$

VI (including III)	VII
$\partial_1 + (x^3 - Ax^2)\partial_2 + (x^2 - Ax^3)\partial_3$	$\partial_1 + (x^3 - Ax^2)\partial_2 - (x^2 + Ax^3)\partial_3$
∂_2	∂_2
∂_3	∂_3
∂_1	∂_1
$e^{-Ax^1}(\cosh x^1\partial_2 + \sinh x^1\partial_3)$	$e^{-Ax^1}(\cos x^1\partial_2 - \sin x^1\partial_3)$
$e^{-Ax^1}(\sinh x^1\partial_2 + \cosh x^1\partial_3)$	$e^{-Ax^1}(\sin x^1\partial_2 + \cos x^1\partial_3)$
dx^1	dx^1
$e^{Ax^1}(\cosh x^1dx^2 - \sinh x^1dx^3)$	$e^{Ax^1}(\cos x^1dx^2 - \sin x^1dx^3)$
$e^{Ax^1}(-\sinh x^1dx^2 + \cosh x^1dx^3)$	$e^{Ax^1}(\sin x^1dx^2 + \cos x^1dx^3)$

VIII
∂_1 $-\sinh x^1 \tanh x^2 \partial_1 + \cosh x^1 \partial_2 - \sinh x^1 \operatorname{sech} x^2 \partial_3$ $\cosh x^1 \tanh x^2 \partial_1 - \sinh x^1 \partial_2 + \cosh x^1 \operatorname{sech} x^2 \partial_3$
$\operatorname{sech} x^2 \cos x^3 \partial_1 - \sin x^3 \partial_2 - \tanh x^2 \cos x^3 \partial_3$ $\operatorname{sech} x^2 \sin x^3 \partial_1 + \cos x^3 \partial_2 - \tanh x^2 \sin x^3 \partial_3$ ∂_3
$\cosh x^2 \cos x^3 dx^1 - \sin x^3 dx^2$ $\cosh x^2 \sin x^3 dx^1 + \sin x^3 dx^2$ $\sinh x^2 dx^1 + dx^3$

IX
∂_1 $\sin x^1 \tan x^2 \partial_1 + \cos x^1 \partial_2 + \sin x^1 \sec x^2 \partial_3$ $\cos x^1 \tan x^2 \partial_1 - \sin x^1 \partial_2 + \cos x^1 \sec x^2 \partial_3$
$\sec x^2 \cos x^3 \partial_1 - \sin x^3 \partial_2 + \tan x^2 \cos x^3 \partial_3$ $\sec x^2 \sin x^3 \partial_1 + \cos x^3 \partial_2 + \tan x^2 \sin x^3 \partial_3$ ∂_3
$\cos x^2 \cos x^3 dx^1 - \sin x^3 dx^2$ $\cos x^2 \sin x^3 dx^1 + \cos x^3 dx^2$ $-\sin x^2 dx^1 + dx^3$

2.5. Isometry groups

In relativity we are interested in Riemannian space-times, and thus in transformations that preserve some or all of the Riemannian structure. For the present purpose we restrict attention to isometries or motions, which are transformations that preserve the metric g. For a continuous group of isometries we will then have, for each generating vector field $\underset{\sim}{Y}$,

$$\mathcal{L}_{\underset{\sim}{Y}} g = 0 \iff Y_{(\mu;\nu)} = 0 \tag{2.16}$$

using (2.3). (2.16) is known as Killing's equation, and its solutions are called Killing vectors (strictly, Killing vector fields). The Killing vectors form the Lie algebra of a Lie group of isometries. The stability group of p within the isometry group is usually called the isotropy group of p; its generators have $\underset{\sim}{Y} = 0$ at p. Consequently it gives rise to a group of linear maps of vectors at p to vectors at p, by (2.1). This is called the linear isotropy group. Since each element of the linear isotropy group preserves the metrix, it must be a subgroup of the appropriate "rotation" group (the Lorentz group for a space-time, or the group SO(3) for a spacelike surface). It can be shown to be isomorphic to the isotropy group, as follows.

Theorem 1: The only Killing vector field which satisfies $\underset{\sim}{Y} = \underset{\sim}{0}$ and $Y_{\mu;\nu} = 0$ at a given point p is the zero field.

Proof: Such a Killing vector field fixes p and any vector at p. Any point q may be connected to p by a geodesic (at least, in the path-connected component containing p). Since $\underset{\sim}{Y}$ preserves the metric it transforms geodesics to geodesics and the length from p to q is also unchanged. Hence the geodesic from p to q, which has a fixed length and given initial tangent vector, does not move and so q is fixed. Thus $\underset{\sim}{Y} = 0$ at q.

Now if two linear isotropies are identical, the Killing vectors $\underset{\sim}{Y}_1$, $\underset{\sim}{Y}_2$ generating them obey $Y_\mu = 0$, $Y_{\mu;\nu} = 0$ at p where $\underset{\sim}{Y} = \underset{\sim}{Y}_1 - \underset{\sim}{Y}_2$, and thus using (2.1) $\underset{\sim}{Y}_1 - \underset{\sim}{Y}_2 = \underset{\sim}{0}$, which implies $\underset{\sim}{Y}_1 = \underset{\sim}{Y}_2$. Hence the isotropies are identical, and so the map of isotropies to linear isotropies is an isomorphism.

Theorem 1 immediately gives another useful result

Theorem 2: If a group of motions G_r of dimension r acts on a Riemannian manifold of dimension m, then

$$r \leqslant m(m+1)/2 \tag{2.17}$$

Proof: Let $\{\underset{\sim}{X}_\alpha\}$ be a basis of Killing vectors of the group. Since

the group is assumed to be effective, $\underset{\sim}{X} = 0$ corresponds uniquely to the identity of G_r. For any constants C^α, $C^\alpha \underset{\sim}{X}_\alpha = \underset{\sim}{Y}$ is Killing. Take any point p, and choose the C^α so that at p, $Y_\mu = 0$ and $Y_{[\mu;\nu]} = 0$. There are $m + m(m-1)/2 = m(m+1)/2$ linear equations for the C^α. If $r > m(m+1)/2$ they have non-zero-solutions, giving $\underset{\sim}{Y} \neq 0$. But Y_μ and $Y_{\mu;\nu}$ are zero at p and so by Theorem 1, $\underset{\sim}{Y} = 0$. Hence $r \leqslant m(m+1)/2$.

The actual value of r is determined by the number of additional conditions to be satisfied by the Y_μ and $Y_{[\mu;\nu]}$ at a point p. This depends on the number of independent conditions arising from the integrability conditions on (2.16). Differentiation yields

$$\mathcal{L}_{\underset{\sim}{Y}} \Gamma = 0 \iff Y_{\mu;\nu\pi} = R_{\mu\nu\pi\varsigma} Y^\varsigma \qquad (2.18)$$

as a first order differential equation for $Y_{\mu;\nu}$, and

$$\mathcal{L}_{\underset{\sim}{Y}}(R_{\mu\nu\pi\varsigma;\sigma_1\sigma_2\cdots\sigma_N}) = 0$$

for N = 0,1,... are the integrability conditions of (2.16) and (2.18).

2.6. Spaces of constant curvature

A Riemannian space is said to be of constant curvature if

$$R_{\mu\nu\pi\varsigma} = K(g_{\mu\pi} g_{\nu\varsigma} - g_{\mu\varsigma} g_{\nu\pi}), \qquad (2.19)$$

where K is a constant. The name arises from considering the sectional curvatures, which are the curvatures of geodesic two-dimensional submanifolds passing through a given point p, defined by drawing all geodesics through p whose initial tangent vector is a linear combination of two given vectors, $\underset{\sim}{u}_1$ and $\underset{\sim}{u}_2$ say. The two-surface has a metric $g_{ab} = g_{\mu\nu} \frac{\partial x^\mu}{\partial y^a} \frac{\partial x^\nu}{\partial y^b}$ where the y^a are some coordinates in the two-surface. The two-surface curvature $R^{(2)}_{abcd}$ has only one independent component, say $R^{(2)}_{1212}$, and its (Gaussian) curvature K is defined as $R^{(2)}_{1212}/\det(g_{ab})$, the sectional curvature being K(p).

Lemma 1: K is independent of the choice of coordinates y^a

Proof: If $y^a \longmapsto y^{a'}$, $R^{(2)\prime}_{1212} = R_{1212}\left(\frac{\partial y^1}{\partial y^{1'}} \frac{\partial y^2}{\partial y^{2'}} - \frac{\partial y^1}{\partial y^{2'}} \frac{\partial y^2}{\partial y^{1'}}\right)^2 = R^{(2)}_{1212} J^2$ where J is the Jacobian determinant of the coordinate change. But $\det(g'_{ab}) = J^2 \det(g_{ab})$, so K is unaltered.

Lemma 2:

$$K(p) = \frac{R_{\mu\nu\pi\varsigma} u_1^\mu u_2^\nu u_1^\pi u_2^\varsigma}{(g_{\mu\pi} g_{\nu\varsigma} - g_{\mu\varsigma} g_{\nu\pi}) u_1^\mu u_2^\nu u_1^\pi u_2^\varsigma} \quad \text{at p}$$

Proof: Let $\underset{\sim}{v}_1$, $\underset{\sim}{v}_2$ be orthogonal unit vectors in the plane of $\underset{\sim}{u}_1$, $\underset{\sim}{u}_2$. Choose coordinates in the 2-surface by assigning the point at distance τ along the geodesic with initial unit tangent vector $\xi^a \underset{\sim}{v}_a$ the coordinates $y^a = \tau \xi^a$. These are called Riemannian normal coordinates. Similarly let $\{x^\alpha\}$ be Riemannian normal coordinates in M. Then the Christoffel symbols of the first kind are

$$\{ab,c\} = v_a^\mu v_b^\nu v_c^\pi \{\mu\nu,\pi\}$$

and at p $\{ab,c\} = 0$. Then

$$R^{(2)}_{1212} = \frac{\partial}{\partial y^1}\{22,1\} - \frac{\partial}{\partial y^2}\{12,1\} \quad \text{at p}$$

$$= v_1^\mu v_2^\nu v_1^\pi v_2^\varsigma R_{\mu\nu\pi\varsigma}$$

To obtain the lemma one now has to remember that the expressions are tensorial, and transform to general coordinates and a general basis $\underset{\sim}{u}_1$, $\underset{\sim}{u}_2$ of the tangent plane to the 2-surface.

In general K depends on p and the plane of $\underset{\sim}{u}_1$ and $\underset{\sim}{u}_2$, but clearly in a space of constant curvature K is independent of these.

If we have a space of constant curvature we can find its metric. This follows from the next set of theorems.

Theorem 3: A Riemannian manifold of dimension $m \geqslant 3$ is of constant curvature if it admits an isotropy group of dimension $m(m-1)/2$ at every point.

Proof: In this case K is independent of the plane of $\underset{\sim}{u}_1$ and $\underset{\sim}{u}_2$ since the isotropy group is isomorphic with the whole of the relevant "rotation" group, and the curvature is invariant under isotropies. Thus $A_{\mu\nu\pi\varsigma} u_1^\mu u_2^\nu u_1^\pi u_2^\varsigma = 0$ for any choice of $\underset{\sim}{u}_1$, $\underset{\sim}{u}_2$, where $A_{\mu\nu\pi\varsigma} = R_{\mu\nu\pi\varsigma} - K(g_{\mu\pi} g_{\nu\varsigma} - g_{\mu\varsigma} g_{\nu\pi})$. By the usual type of "tensor detection theorem" or "quotient theorem" argument

$$A_{\mu\nu\pi\varsigma} + A_{\mu\varsigma\pi\nu} + A_{\pi\nu\mu\varsigma} + A_{\pi\varsigma\mu\nu} = 0$$

and hence, since $A_{\mu\nu\pi\varsigma}$ has the symmetries of $R_{\mu\nu\pi\varsigma}$, we can deduce that $A_{\mu\nu\pi\varsigma} = 0$. Finally the Bianchi identities $R_{\mu\nu[\pi\varsigma;\sigma]} = 0$ contracted on $\nu\varsigma$ give

$$(m-2)(K_{,\mu} g_{\sigma\pi} - K_{,\sigma} g_{\mu\pi}) = 0$$

and thus if $m \geqslant 3$ we can contract again to give

$$(m-1)K_{,\mu} = 0 \qquad K \text{ is constant.}$$

Theorem 4: A Riemannian manifold of dimension m is of constant curva-

ture if and only if it admits an isometry group of dimension $m(m+1)/2$.

Proof: "if" : By (2.15) $s = m(m-1)/2$. Then for $m \geqslant 3$, Theorem 1 proves the result. For $m = 2$, K is the only independent curvature component and since the whole manifold is a single orbit K must be constant.

"only if". The conditions $\mathcal{L}_\xi R_{\mu\nu\pi\rho;\sigma_1\sigma_2\ldots} = 0$ are all identically satisfied. Hence by the argument at the end of Section 2.5 the result is proved.

Corollary 1: The converse of Theorem 4 is true.

Proof: using "only if" part of Theorem 4 and (2.15).

Corollary 2: A two-dimensional space admitting a G_2 of motions admits a G_3 of motions.

Proof: The G_2 must be transitive, to satisfy (2.15). Thus K is the same at all points. Thus by Theorem 4, the space admits a G_3.

Theorem 5: Any two metrics of the same constant curvature K and the same signature are equivalent.

Proof [20]: Starting from any set of Riemannian normal coordinates $\{y^\mu\}$ one can transform to a set

$$x^\mu = \sum_{K=1} \frac{1}{K!} \left(\frac{\partial^K x^\mu}{\partial y^{\nu_1} \partial y^{\nu_2} \ldots \partial y^{\nu_K}} \right) y^{\nu_1} \ldots y^{\nu_K}$$

where $\dfrac{\partial x^\mu}{\partial y^\nu} = \delta^\mu{}_\nu$, $\dfrac{\partial^2 x^\mu}{\partial y^{\nu_1} \partial y^{\nu_2}} = -\Gamma^\mu{}_{\nu_1\nu_2}$ etc.

at the origin, so that there

$$\partial_{(\lambda_1 \ldots \lambda_r} \Gamma^\mu{}_{\nu_1\nu_2)} = 0$$

for $r = 0, 1, \ldots$, in the $\{x^\mu\}$ system. In this system the metric tensor is

$$g_{\mu\nu} = g^{(0)}_{\mu\nu} - \frac{1}{3} R^{(0)}_{\mu\eta\nu\lambda} y^\eta y^\lambda - \frac{1}{3!} R^{(0)}_{\mu\eta\nu\lambda;\rho} y^\eta y^\lambda y^\rho$$

superscript o denoting the value at the origin, and the metric is thus fixed by $R^{(0)}_{\mu\eta\nu\lambda}$ in our case.

Finally one simply has to construct a particular metric of constant curvature, and this is done by guessing, it is of the form $e^{2\sigma}(d\hat{s}^2)$ where $d\hat{s}^2$ is a flat space of the right signature, and computing the necessary form for e^σ. The result is

$$e^{-\sigma} = \sqrt{|K|}\,(1 + \frac{k}{4} \sum_\mu l_\mu (x^\mu)^2) \qquad (2.20)$$

where $k = \mathrm{sgn}(K)$, the l_μ are $+1$ or -1 in accordance with the required signature, and $d\hat{s}^2 = \sum_\mu l_\mu (dx^\mu)^2$.

The results of this section are all well-known classical theorems and appear in many texts, e.g. [8], [20].

2.7. Metrics with isometries

If the manifold admits a simply-transitive group of motions it is diffeomorphic to the group itself. Using the basis of vector fields $\underset{\sim}{W}_\alpha$, so that $g = g_{\alpha\beta}\omega^\alpha\omega^\beta$ and $g_{\alpha\beta} = g(\underset{\sim}{W}_\alpha, \underset{\sim}{W}_\beta)$ we find, treating $g_{\alpha\beta}$ as a scalar,

$$\begin{aligned}\mathcal{L}_{\underset{\sim}{Y}} g_{\alpha\beta} &= \mathcal{L}_{\underset{\sim}{Y}}(g(\underset{\sim}{W}_\alpha, \underset{\sim}{W}_\beta)) = \\ &= (\mathcal{L}_{\underset{\sim}{Y}} g)(\underset{\sim}{W}_\alpha, \underset{\sim}{W}_\beta) + g(\mathcal{L}_{\underset{\sim}{Y}}\underset{\sim}{W}_\alpha, \underset{\sim}{W}_\beta) + g(\underset{\sim}{W}_\alpha, \mathcal{L}_{\underset{\sim}{Y}}\underset{\sim}{W}_\beta) = 0\end{aligned}$$

so the $g_{\alpha\beta}$ are constants. Note that the reciprocal group is in general not an isometry group.

When the group acts simply-transitively on orbits which are submanifolds, it is possible at any point p of an orbit, to complete $\{\underset{\sim}{W}_\alpha\}$ $\alpha = 1,\ldots,m$ to be a basis $\{\underset{\sim}{W}_\mu\}$ $\mu = 1,\ldots,n$ of $\mathcal{T}_p(M)$ by adding additional vectors, and a basis at other points of $\{pT_a;\ a \in G\}$ can then be defined by the condition $\mathcal{L}_{\underset{\sim}{Y}}\underset{\sim}{W}_\mu = 0$ for $\mu = m+1,\ldots,n$. By the argument above, $g_{\mu\nu}$ in this basis is constant within each orbit.

If the orbits of the group are hypersurfaces (i.e. submanifolds of dimension $m = n - 1$), we can prove that the (unit) normals of those hypersurfaces are geodesic. Denote the normal by $\underset{\sim}{n}$. Then taking any Killing vector basis $\{\underset{\sim}{X}_\alpha\}$ we must have $g(\underset{\sim}{n}, \underset{\sim}{X}_\alpha) = 0$ at all points. Hence

$$0 = \mathcal{L}_{\underset{\sim}{X}_\alpha}(g(\underset{\sim}{n}, \underset{\sim}{X}_\beta)) = (\mathcal{L}_{\underset{\sim}{X}_\alpha} g)(\underset{\sim}{n}, \underset{\sim}{X}_\beta) + g(\mathcal{L}_{\underset{\sim}{X}_\alpha}\underset{\sim}{n}, \underset{\sim}{X}_\beta) + g(\underset{\sim}{n}, \mathcal{L}_{\underset{\sim}{X}_\alpha}\underset{\sim}{X}_\beta)$$

and since $\mathcal{L}_{\underset{\sim}{X}_\alpha} g = 0$, $\mathcal{L}_{\underset{\sim}{X}_\alpha}\underset{\sim}{X}_\beta = C^\gamma{}_{\alpha\beta}\underset{\sim}{X}_\gamma$, and the $\{\underset{\sim}{X}_\beta\}$ at any point span $\mathcal{T}_p$ (orbit), we find $\mathcal{L}_{\underset{\sim}{X}_\alpha}\underset{\sim}{n}$ is normal to the orbit. But $g(\underset{\sim}{n}, \underset{\sim}{n})$ is constant, so $\mathcal{L}_{\underset{\sim}{X}_\alpha}\underset{\sim}{n} = 0$, and $n^\mu n_{\mu;\nu} n^\nu = 0$. Finally

$$0 = \mathcal{L}_{\underset{\sim}{n}}(g(\underset{\sim}{n}, \underset{\sim}{X}_\beta)) = n_{\mu;\nu} n^\nu X_\beta{}^\mu + (X_\beta)_{\mu;\nu} n^\nu n^\mu = n_{\mu;\nu} n^\nu X_\beta{}^\mu$$

using (2.16), and thus $n^\mu{}_{;\nu} n^\nu = 0$. We may now take the coordinate t to be the affine parameter along these geodesics and obtain, if $\underset{\sim}{n}$ is non-null,

$$ds^2 = \varepsilon\, dt^2 + g_{\mu\nu}\, dx^\mu dx^\nu \qquad (2.21)$$

where x^μ; $\mu = 1, 2, 3, \ldots, (n-1)$ are coordinates in the orbits, and $\varepsilon = \pm 1$ depending on the nature of n.

It is easy to prove the existance of manifolds of dimension n with submanifolds of dimension m on which a group acts simply-transitively: one has only to take any metric $g_{\alpha\beta}\omega^\alpha\omega^\beta$ with ω^α defined so that $\mathcal{L}_{\underset{\sim}{Y}}\omega^\alpha = 0$ on each orbit and $g_{\alpha\beta}$ depend only on coordinates labelling different orbits. Schmidt [9] has also proved the existance of manifolds admitting Lie groups with isometries specified by arbitrary Lie algebras subject to certain conditions. These questions are not of significance for our purpose here.

In the same paper [9] and in his thesis [10], Schmidt gave a method for computing possible Lie algebras of Killing vectors multiply-transitive on orbits of a given signature. The idea is that one can choose a basis of Killing vectors $\{\underset{\sim}{X}_\alpha\}$ in such a way that at a chosen point p, $\underset{\sim}{X}_\alpha = 0$ for $\alpha = m+1, \ldots, r$. Since these $\underset{\sim}{X}_\alpha$ correspond to the generators of the linear isotropy group, which is itself a subgroup of the appropriate "rotation" group, one can specify the subgroup to be investigated and then choose the $\underset{\sim}{X}_\alpha$, $\alpha = m+1, \ldots, r$, to give a standard canonical form of the commutation relations for this group. Let us denote such $\underset{\sim}{X}_\alpha$ by $\underset{\sim}{Y}_i$, $i = 1, \ldots, s$. Then

$$\left[\underset{\sim}{Y}_i, \underset{\sim}{Y}_j\right] = f_{ij}^K \underset{\sim}{Y}_K$$

where the f_{ij}^K are known. Next one can choose the Killing vectors $\underset{\sim}{X}_\alpha$, $\alpha = 1, \ldots, m$, which are non-zero at p, to be an orthonormal basis of τ_p (orbit) and further adjust this choice to the choice of $\underset{\sim}{Y}_i$ to give a simple form for the commutators

$$\left[\underset{\sim}{Y}_i, \underset{\sim}{X}_\alpha\right] = \mathcal{L}_{\underset{\sim}{Y}_i}(\underset{\sim}{X}_\alpha), \quad i = 1, \ldots, s, \quad \alpha = 1, \ldots, m \qquad (2.22)$$

at p. The action at p of the $\underset{\sim}{Y}_i$ on the basis $\underset{\sim}{X}_\alpha$ is fixed by the properties of the orthogonal group, so the coefficients of the $\underset{\sim}{X}_\alpha$ in the commutator (2.22) are known. The remaining unknowns are the structure constants in $\left[\underset{\sim}{X}_\alpha, \underset{\sim}{X}_\beta\right]$ ($\alpha, \beta = 1, \ldots, m$) and the coefficients of the $\underset{\sim}{Y}_i$ in (2.22), which are so far undetermined because $\underset{\sim}{Y}_i = 0$ at p. These remaining constants must satisfy the Jacobi identities, and thus all possibilities can be enumerated.

This method has the further advantage (not explored by Schmidt) that if the orbit has a simply-transitive group contained in the multiply-transitive group, it is very easy to find because it must have a Killing vector basis $\{\underset{\sim}{Z}_\alpha\}$ $\alpha = 1, \ldots, m$ equal at p to $\{\underset{\sim}{X}_\alpha\}$ $\alpha = 1, \ldots m$. The two bases must obey

$$\underset{\sim}{Z}_\alpha = \underset{\sim}{X}_\alpha + A^i{}_\alpha \underset{\sim}{Y}_i$$

for some constants $A^i{}_\alpha$ and one can find all simply-transitive subgroups by taking arbitrary $A^i{}_\alpha$ and imposing the condition that the $\underset{\sim}{Z}_\alpha$ form the basis of a subalgebra.

The above description may be rather hard to follow, so I shall give an example relevant to the next chapter.

Consider a positive-definite two-dimensional manifold of constant curvature K. We know (by Theorem 4) that this admits a G_3 of motions and hence that s = 1, from (2.15). This isotropy group is the rotation group of a plane, so we can choose $\underset{\sim}{X}_1$, $\underset{\sim}{X}_2$, $\underset{\sim}{Y}$ so that at p

$$\begin{aligned} [\underset{\sim}{Y},\underset{\sim}{X}_1] &= \underset{\sim}{X}_2 \\ [\underset{\sim}{Y},\underset{\sim}{X}_2] &= -\underset{\sim}{X}_1 \end{aligned} \tag{2.23}$$

Thus the full set of commutators must be

$$\begin{aligned} [\underset{\sim}{Y},\underset{\sim}{X}_1] &= \underset{\sim}{X}_2 + a\underset{\sim}{Y} \\ [\underset{\sim}{Y},\underset{\sim}{X}_2] &= -\underset{\sim}{X}_1 + b\underset{\sim}{Y} \\ [\underset{\sim}{X}_1,\underset{\sim}{X}_2] &= A\underset{\sim}{X}_1 + B\underset{\sim}{X}_2 + C\underset{\sim}{Y} \end{aligned} \tag{2.24}$$

Moreover, $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$ are only fixed by (2.23) up to the addition of a term in $\underset{\sim}{Y}$, so by taking a new basis $\underset{\sim}{X}_1' = \underset{\sim}{X}_1 - b\underset{\sim}{Y}$, $\underset{\sim}{X}_2' = \underset{\sim}{X}_2 + a\underset{\sim}{Y}$ we can eliminate a and b. Taking these as zero, we can apply the Jacobi identity (2.4) to ($\underset{\sim}{X}_1$, $\underset{\sim}{X}_2$, $\underset{\sim}{Y}$) to obtain $A\underset{\sim}{X}_2 - B\underset{\sim}{X}_1 = 0$. This implies A=B=0 and there are three possibilities, $C > 0$, $C = 0$ and $C < 0$. By scaling $\underset{\sim}{X}_1$, $\underset{\sim}{X}_2$ and $\underset{\sim}{Y}$ we could set C to ± 1 or 0, but $\underset{\sim}{X}_1$, $\underset{\sim}{X}_2$ would not then be orthonormal. Schmidt [9] has shown how to calculate the Riemann tensor of the orbit of a multiply-transitive group from its Lie algebraic structure, and it can be verified that C = K. The Bianchi types of (2.24) are VIII ($K < 0$), VIII_0 ($K = 0$) or IX ($K > 0$). Now on taking a new basis $\underset{\sim}{Z}_1 = \underset{\sim}{X}_1 + \alpha\underset{\sim}{Y}$, $\underset{\sim}{Z}_2 = \underset{\sim}{X}_2 + \beta\underset{\sim}{Y}$, we find

$$[\underset{\sim}{Z}_1,\underset{\sim}{Z}_2] = -\alpha\underset{\sim}{Z}_1 - \beta\underset{\sim}{Z}_2 + (C + \alpha^2 + \beta^2)\underset{\sim}{Y}$$

so that there is a different simply-transitive subgroup corresponding to every possible choice of α and β satisfying

$$C + \alpha^2 + \beta^2 = 0$$

When $C > 0$ this has no solutions. When $C = 0$ there is exactly one solution, $\alpha = \beta = 0$, and the simply-transitive group is of type G_2I.

When $C < 0$, there is a one-parameter family of simply-transitive groups of type G_2II arising from the solutions $\alpha = |C|\cos\phi$, $\beta = |C|\sin\phi$ for arbitrary ϕ .

The advantage of finding a simply-transitive subgroup is that one can use its reciprocal group to put the metric in a simple form, and so make computations of curvature more quickly than by Schmidt's method. In fact the reciprocal group basis can be chosen so that it is orthonormal everywhere, and agrees with $\underset{\sim}{Z}_\alpha$ at p.

Theorem 6: A Riemannian manifold of dimension $m > 2$ cannot admit a maximal group of motions of dimension $m(m+1)/2 - 1$.

Proof: We first prove the result for $m = 3$. Here $m(m+1)/2 - 1 = 5$ and the isotropy group of a point p has dimension 2. It therefore acts on 2-dimensional submanifolds (hypersurfaces), and thus the metric has the form (2.21). The hypersurfaces have curvature K(t) and their metric can be written in the form (2.20). Then the third Killing vector of (2.20) for a given t clearly is Killing for all t and preserves (2.21). Thus p has an isotropy group with $s = 3$, contrary to hypothesis.

Now for $m > 3$ we prove the result by induction. Using an analogous argument to that above we see that there are $(m-1)$ dimensional orbits of a group of dimension $m(m+1)/2 - 1 - m = m(m-1)/2$ contrary to the hypothesis for dimension $(m-1)$. But the induction starts at $m = 3$, and is thus valid for all $m > 2$.

3. Spatially-homogeneous cosmological metrics

3.1. Bianchi models

A space-time is said to be spatially-homogeneous if it admits a group G_r of isometries acting transitively on spacelike hypersurfaces. If we put $m = 3$ in (2.15) and Theorems 2 and 6 we see that the only possibilities are: $r = 3$, $s = 0$; $r = 4$, $s = 1$; and $r = 6$, $s = 3$. We take each of these in turn, starting with $r = 3$.

Spatially-homogeneous cosmologies with $r = 3$ are known as Bianchi models, because they can be classified according to the Bianchi type of their G_3 of motions. The metrics can easily be written down (in terms of unknown functions of time) in any of a number of different forms. As described in Section 2.4, one can take any point p of M, and assign coordinates on the orbit through p in accordance with Table III. To assign the coordinates on other orbits one can choose any vector $\underset{\sim}{y}$ at p not lying in the orbit, and any curve $\gamma(y)$ with $\underset{\sim}{y}$ as

its initial tangent vector, and then propagate this vector and curve by dragging along by the Killing vectors. The coordinates then be made co-moving with respect to y. The obvious choice for y is the t of (2.21). If one introduces the reciprocal group bases given in Table III, one will have $\left[\frac{\partial}{\partial t}, \underset{\sim}{W}_a\right] = 0$, where a = 1,2,3 hence the whole metric is

$$ds^2 = - dt^2 + g_{ab}(t)\, \omega^a \omega^b \qquad (3.1)$$

g_{ab} being a function of t alone. This form of the metric is used extensively in the literature (see e.g. Ryan and Shepley [11]).

A slightly different choice arises in a natural way when the matter content specifies uniquely a (unit) vector field $\underset{\sim}{u}$ other than $\underset{\sim}{n} = \frac{\partial}{\partial t}$, such as when the matter content is a perfect fluid flowing with four velocity $\underset{\sim}{u} \neq \underset{\sim}{n}$. Then one can take $\underset{\sim}{u} = \frac{\partial}{\partial t}$, say, and make the space coordinates co-moving with respect to τ. Introducing the $\underset{\sim}{W}_a$ as before we will obtain

$$ds^2 = - d\tau^2 + 2(g_{ab}u^b)\, \omega^a d\tau + g_{ab}\, \omega^a \omega^b \qquad (3.2)$$

where the g_{ab} are the same as in (3.1), and $\underset{\sim}{u} = u^4 \underset{\sim}{n} + u^a \underset{\sim}{W}_a$. (Note that since the $\underset{\sim}{W}_a$ commute with the $\underset{\sim}{X}_a$, as do $\underset{\sim}{u}$ and $\underset{\sim}{n}$, the change of coordinate origin in each orbit will not alter the actual fields $\underset{\sim}{W}_a$ although it will alter their coordinate form at a given point of M). This is easily derived by finding the dual basis to $\{\underset{\sim}{u}, \underset{\sim}{W}_a\}$ in terms of the dual basis of $\{\underset{\sim}{n}, \underset{\sim}{W}_a\}$; it is $\{d\tau, \hat{\omega}^a\}$ where $dt = u^4 d\tau$, $\hat{\omega}^a = u^a d\tau + \omega^a$; and this can then be substituted in (3.1).

A quite different way of writing the metrics is to choose an orthonormal tetrad of vectors $\{\underset{\sim}{e}_\alpha;\ \alpha = 1,\ldots,4\}$ in each orbit in such a way that $\underset{\sim}{e}_4 = \underset{\sim}{n}$ and the $\underset{\sim}{e}_a$ are reciprocal group generators. This possibility has been exploited by Ellis and MacCallum [4], and, in a different notation by Estabrook et al. [5]. If the dual basis is $\{\underset{\sim}{e}^\alpha,\ \alpha = 1\ldots4\}$, the metric is of course

$$ds^2 = -(\underset{\sim}{e}^4)^2 + (\underset{\sim}{e}^1)^2 + (\underset{\sim}{e}^2)^2 + (\underset{\sim}{e}^3)^2 \qquad (3.3)$$

Since $\underset{\sim}{e}^4$ is exact (being just dt), $d\underset{\sim}{e}^4 = 0$. The commutators of the $\{\underset{\sim}{e}_a,\ a = 1,2,3\}$ must have the same Lie algebra structure as the G_3 itself, and the method of classification given in Section 2.4 can be applied using the metric induced by that of space-time, $g_{ab} = \delta^a_b$. However, it is not permissible to scale the $\underset{\sim}{e}_a$ (or the orthonormality is lost) so their commutators can at best be reduced to a form

$$[\underset{\sim}{e}_a, \underset{\sim}{e}_b] = \gamma^c{}_{ab} \underset{\sim}{e}_c \qquad (3.4)$$

with

$$\frac{1}{2}\,\gamma^{c}{}_{ab}\,\varepsilon^{abd} = n^{cd} + \varepsilon^{cde}\,a_e \tag{3.5}$$

and

$$a_e = (a,0,0), \qquad n^{cd} = \begin{pmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{pmatrix} \tag{3.6}$$

where h $n_2\, n_3 = a^2$ and a, n_1, n_2, n_3 are functions of t alone. The relation of this choice to (3.1) is rather awkward. In a given orbit the $\underset{\sim}{e}_b$ are proportional to one of the (a - d) dimensional set of choices of $\underset{\sim}{W}_b$ consistent with Table I, but since $|e_4, W_b| = 0$ the condition that this remains true in successive orbits, with a fixed $\underset{\sim}{W}_b$ as used in (3.1), is that

$$[\underset{\sim}{e}_4, \underset{\sim}{e}_b] = -\Theta_b \underset{\sim}{e}_b \qquad \text{(no sum on b)}. \tag{3.7}$$

It turns out that certain results are much easier to obtain using the form (3.3) - (3.6) than with (3.1), and vice versa. There are other results for which yet another formalisms have proved helpful. One of these is to adopt (3.3) but with $\underset{\sim}{e}_4 \neq \underset{\sim}{n}$. This was used by Ellis and King [12, 13] to investigate cosmologies with fluid flowing with velocity $\underset{\sim}{u} \neq \underset{\sim}{n}$ (called "titled" models). They also introduced a non-orthonormal tetrad $\{\underset{\sim}{u}, \underset{\sim}{e}_a\}$ with the $\underset{\sim}{e}_a$ being the reciprocal group generators obeying (3.4), (3.5). Yet another formalism has been exploited by Siklos [7, 14]. This is to take a null tetrad $\{\underset{\sim}{l}, \underset{\sim}{k}, \underset{\sim}{m}, \underset{\sim}{\bar{m}}\}$ as used in the well-known Newman-Penrose formalism [15], with the tetrad chosen so that $\underset{\sim}{l}$, $\underset{\sim}{k}$ and $\underset{\sim}{n}$ are coplanar and the tetrad is invariant under the group G_3. This formalism is useful for studying those spatially-homogeneous models which have algebraically special Weyl tensors [14] or horizons where the spacelike hypersurfaces become null [7].

3.2. Spatially-homogeneous cosmologies with a rotational symmetry

This section is concerned with space-times admitting a G_4 transitive on spacelike hypersurfaces. It has been proved by Kantowski [16] that there is only one case in which the G_4 does not contain a simply-transitive G_3. A more accessible version of Kantowski's proof has been given by Collins [17]. The exceptional case is known as the Kantowski-Sachs metric, since it appeared in [18], being Case I in that paper, but, as Collins pointed out, it had been used before, no-

tably by Kompaneets and Chernov [19]. However, Kompaneets and Chernov did not recognize its exceptional character, nor, indeed, did the Kantowski-Sachs paper give a quite correct account since they thought the property was shared by another metric (Case II of their paper). The reason for this error was that they were unaware that the G_3 of a two-dimensional space of constant negative curvature contains (as I showed in Section 2.7) a simply-transitive G_2. This was significant because Kantowski's proof consisted of first proving Egorov's theorem (as stated, but not proved, in Petrov [20]) that every G_4 contains a G_3, then noting that if this G_3 was not simply-transitive it acted on two-dimensional spaces of constant curvature, treating each of the algebras (2.24) by adding a fourth Killing vector and studying the Jacobi identities, and finally seeking simply-transitive subgroups of the resulting groups.

Using Schmidt's ideas, as outlined in Section 2.7, I will now give a new and different proof, which will simultaneously reveal the structure of the other space-time with a G_4 transitive on spacelike hypersurfaces.

We take the isotropy $\underset{\sim}{Y}$ as a spatial rotation with axis $\underset{\sim}{X}_1$ at p, $\underset{\sim}{X}_1$, $\underset{\sim}{X}_2$, $\underset{\sim}{X}_3$ being an orthonormal basis of non-zero Killing vectors at p. Then

$$\begin{aligned}
[\underset{\sim}{Y},\underset{\sim}{X}_1] &= C_1\underset{\sim}{Y} \\
[\underset{\sim}{Y},\underset{\sim}{X}_2] &= \underset{\sim}{X}_3 + C_2\underset{\sim}{Y} \\
[\underset{\sim}{Y},\underset{\sim}{X}_3] &= -\underset{\sim}{X}_2 + C_3\underset{\sim}{Y}
\end{aligned}$$

By a change of basis to $\underset{\sim}{X}_2' = \underset{\sim}{X}_2 - C_3\underset{\sim}{Y}$, $\underset{\sim}{X}_3' = \underset{\sim}{X}_3 + C_2\underset{\sim}{Y}$ we can eliminate C_2 and C_3, so we take those to be zero. We now write

$$\begin{aligned}
[\underset{\sim}{X}_2,\underset{\sim}{X}_3] &= N_1\underset{\sim}{X}_1 + (N_{12} - A_3)\underset{\sim}{X}_2 + (N_{13} + A_2)\underset{\sim}{X}_3 + B_{23}\underset{\sim}{X}_4 \\
[\underset{\sim}{X}_3,\underset{\sim}{X}_1] &= (N_{12} + A_3)\underset{\sim}{X}_1 + N_2\underset{\sim}{X}_2 + (N_{23} - A_1)\underset{\sim}{X}_3 + B_{31}\underset{\sim}{X}_4 \\
[\underset{\sim}{X}_1,\underset{\sim}{X}_2] &= (N_{13} - A_2)\underset{\sim}{X}_1 + (N_{23} + A_1)\underset{\sim}{X}_2 + N_3\underset{\sim}{X}_3 + B_{12}\underset{\sim}{X}_4
\end{aligned}$$

and apply the Jacobi identities. The Jacobi identities for $(\underset{\sim}{X}_3,\underset{\sim}{X}_1,\underset{\sim}{Y})$ and $(\underset{\sim}{X}_1,\underset{\sim}{X}_2,\underset{\sim}{Y})$ yield

$$\begin{aligned}
&(N_{13} - A_2)\underset{\sim}{X}_1 + (2N_{23} - C_1)\underset{\sim}{X}_2 + (N_3 - N_2)\underset{\sim}{X}_3 + (B_{12} + A\,(N_{12} + A_3))\underset{\sim}{Y} = 0 \\
&-(N_{12} + A_3)\underset{\sim}{X}_1 - (N_3 - N_2)\underset{\sim}{X}_2 - (2N_{23} + C_1)\underset{\sim}{X}_3 - (B_{31} - A(N_{13} - A_2))\underset{\sim}{Y} = 0
\end{aligned}$$

which imply

$$N_{23} = C_1 = B_{12} = B_{31} = 0 = N_{13} - A_2 = N_{12} + A_3 = N_2 - N_3$$

Using these, the Jacobi identity for $(\underset{\sim}{X}_2, \underset{\sim}{X}_3, \underset{\sim}{Y})$ gives

$$(N_{13} + A_2)\underset{\sim}{X}_2 - (N_{12} - A_3)\underset{\sim}{X}_3 = 0$$

so $N_{13} = A_2 = N_{12} = A_3 = 0$, and lastly the Jacobi identities for $(\underset{\sim}{X}_1, \underset{\sim}{X}_2, \underset{\sim}{X}_3)$ yield

$$2A_1N_1\underset{\sim}{X}_1 + 2A_1B_{23}\underset{\sim}{Y} = 0$$

so that $A_1N_1 = A_1B_{23} = 0$. Thus we have proved the following

Lemma 3: Any group G_4 acting on a positive-definite three-dimensional manifold has a basis of Killing vectors such that the commutators are

$$\begin{aligned}
&[\underset{\sim}{Y},\underset{\sim}{X}_1] = 0, \quad [\underset{\sim}{Y},\underset{\sim}{X}_2] = \underset{\sim}{X}_3, \quad [\underset{\sim}{Y},\underset{\sim}{X}_3] = -\underset{\sim}{X}_2 \\
&[\underset{\sim}{X}_2,\underset{\sim}{X}_3] = N_1\underset{\sim}{X}_1 + B\underset{\sim}{Y}, \quad [\underset{\sim}{X}_3,\underset{\sim}{X}_1] = N_2\underset{\sim}{X}_2 - A\underset{\sim}{X}_3 \qquad (3.7)\\
&[\underset{\sim}{X}_1,\underset{\sim}{X}_2] = A\underset{\sim}{X}_2 + N_2\underset{\sim}{X}_3
\end{aligned}$$

where N_1, N_2, A, B are constants subject to $N_1A = AB = 0$.

These algebras were found by Schmidt [10] under the additional assumption that $C_1 = 0$. As we have seen, this assumption is not necessary.

Next we seek the simply transitive subgroups. Setting $\underset{\sim}{Z}_1 = \underset{\sim}{X}_1 + \alpha\underset{\sim}{Y}$, $\underset{\sim}{Z}_2 = \underset{\sim}{X}_2 + \beta\underset{\sim}{Y}$, $\underset{\sim}{Z}_3 = \underset{\sim}{X}_3 + \gamma\underset{\sim}{Y}$ we obtain

$$\begin{aligned}
[\underset{\sim}{Z}_2,\underset{\sim}{Z}_3] &= N_1\underset{\sim}{Z}_1 - \beta\underset{\sim}{Z}_2 - \gamma\underset{\sim}{Z}_3 + (B - \alpha N_1 + \beta^2 + \gamma^2)\underset{\sim}{Y}\\
[\underset{\sim}{Z}_3,\underset{\sim}{Z}_1] &= (N_2 + \alpha)\underset{\sim}{Z}_2 - A\underset{\sim}{Z}_3 - (\beta(N_2 + \alpha) - A\gamma)\underset{\sim}{Y}\\
[\underset{\sim}{Z}_1,\underset{\sim}{Z}_2] &= A\underset{\sim}{Z}_2 + (N_2 + \alpha)\underset{\sim}{Z}_3 - (A\beta + (N_2 + \alpha)\gamma)\underset{\sim}{Y}
\end{aligned}$$

The coefficients of $\underset{\sim}{Y}$ in the last two equations vanish only if either

$$N_2 + \alpha = A = 0 \quad \text{or} \quad \beta = \gamma = 0 \qquad (3.8)$$

If $\beta = \gamma = 0$, the coefficient of $\underset{\sim}{Y}$ in $[\underset{\sim}{Z}_2,\underset{\sim}{Z}_3]$ vanishes only if $B = \alpha N_1$. Thus if $N_1 \neq 0$ there is a unique α, and if $N_1 = 0$ there is no value of α unless $B = 0$ also, in which case α is arbitrary.

If $N_2 + \alpha = A = 0$, the coefficient of $\underset{\sim}{Y}$ in $[\underset{\sim}{Z}_2,\underset{\sim}{Z}_3]$ vanishes only if

$$B + N_1N_2 + \beta^2 + \gamma^2 = 0 \qquad (3.9)$$

For $B + N_1N_2 > 0$ there is no solution and for $B + N_1N_2 = 0$ only the solution $\beta = \gamma = 0$. If $B + N_1N_2 < 0$ there is a one-parameter family of solutions $\beta = k\cos\phi$, $\gamma = k\sin\phi$ where $k^2 = -(B + N_1N_2)$ and ϕ is arbitrary.

These results show that the only case with no simply-transitive subgroup has $N_1 = 0$, $B > 0$, for then neither of the two possibilities (3.8) yields a simply-transitive group, although the first can be used to set $N_2 = 0$ in (3.7) without affecting the other commutators. Doing this last, we find

$$[\underset{\sim}{X}_2, \underset{\sim}{X}_3] = B\underset{\sim}{Y}, \quad [\underset{\sim}{X}_3, \underset{\sim}{X}_1] = 0 = [\underset{\sim}{X}_1, \underset{\sim}{X}_2]$$

and ($\underset{\sim}{X}_2$, $\underset{\sim}{X}_3$, $\underset{\sim}{Y}$) are the generators of a multiply-transitive Bianchi type IX G_3 (because they give a subalgebra one of whose generators is zero at p) acting on two-dimensional surfaces of constant positive curvature B. This group cannot be multiply-transitive at some points and simply-transitive at others because the commutators imply that for $\underset{\sim}{v}_\alpha = \{\underset{\sim}{X}_2, \underset{\sim}{X}_3, \text{ or } \underset{\sim}{Y}\}$, $\mathcal{L}_{\underset{\sim}{X}_1} \underset{\sim}{v}_\alpha = 0 = \mathcal{L}_{\underset{\sim}{v}_\alpha} \underset{\sim}{X}_1$ and hence $g(\underset{\sim}{v}_\alpha, \underset{\sim}{X}_1) = 0$ at all points, being zero at p. Thus the $\underset{\sim}{v}_\alpha$ always lie in a plane perpendicular to $\underset{\sim}{X}_1$, which is itself of constant length 1, and therefore they can never generate a three-dimensional orbit.

Up to now we have considered only one orbit in the space-time but we can easily find the four-dimensional metric. We define t as in (2.21). Then $[\underset{\sim}{n}, \underset{\sim}{Y}] = 0$ implies that on the integral curve of $\underset{\sim}{n}$ through p, $\underset{\sim}{Y}$ is always zero, and the commutators show that $\underset{\sim}{X}_1$ is always the axis of rotation (along this integral curve) being the only fixed vector in the orbits, and that $\underset{\sim}{X}_2$, $\underset{\sim}{X}_3$ always lie in the plane of rotation and are of equal length. Thus all that can change from one orbit to the next are the lengths of $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$. Since $[\underset{\sim}{X}_1, \underset{\sim}{X}_2] = 0$ we may introduce a coordinate x such that $\underset{\sim}{X}_1 = \frac{\partial}{\partial x}$, and the surface generated by $\underset{\sim}{X}_2$, $\underset{\sim}{X}_3$ may be written as

$$\frac{dy^2 + dz^2}{B(1 + \frac{1}{4}(y^2 + z^2))^2}$$

(using (2.20) with k = 1). The metric has $\underset{\sim}{X}_1$ orthogonal to $\underset{\sim}{n}$, $\underset{\sim}{X}_2$, $\underset{\sim}{X}_3$ and B dependent only on t, (since $\frac{\partial}{\partial x}$ is a Killing vector). Thus the full metric can be written

$$ds^2 = -dt^2 + X^2(t)dx^2 + Y^2(t)(d\theta^2 + \sin^2\theta d\phi^2) \qquad (3.10)$$

where $\tan\phi = z/y$, $r = \sqrt{y^2 + z^2} = z\tan\theta/2$, and X and Y are functions of t alone.

The remaining cases contained in (3.7) all have simply-transitive subgroups, and by an argument similar to that in the Kantowski-Sachs case, $\underset{\sim}{Y} = 0$ all along the integral curve of n through p and $\underset{\sim}{Z}_1$, $\underset{\sim}{Z}_2$, $\underset{\sim}{Z}_3$ are always orthogonal along that curve. Hence the metric,

in terms of reciprocal group generators agreeing with $\underset{\sim}{Z}_\alpha$ at every point along the integral curve of $\underset{\sim}{n}$ through p, must be

$$ds^2 = -dt^2 + X^2(t)\,\omega^1\omega^1 + Y^2(t)(\omega^2\omega^2 + \omega^3\omega^3) \qquad (3.11)$$

the ω^α being, of course, independent of t.

The possible cases that can arise are as follows. We exploit the remaining freedom of choice of $\underset{\sim}{X}_2$, $\underset{\sim}{X}_3$ to set $N_1 > 0$ if $N_1 \neq 0$.

1. If $N_1 \neq 0$ we use α to set $B = 0$. Then there are three possibilities
   a) $N_2 = 0$. In this case (3.9) allows no further possible simply-transitive subgroups and the only one is of Bianchi type II
   b) $N_2 > 0$. Again (3.9) allows no other simply-transitive subgroups, and the one found is of Bianchi type IX
   c) $N_2 < 0$. This gives a group of Bianchi type VIII. Taking $\alpha = -N_2$, β and γ non-zero gives a one-parameter family of simply-transitive groups of type III. For these groups an orthonormal basis of reciprocal group generators at p will have $n^a{}_a = N_1 \neq 0$.

These three metrics have a form of the Taub-NUT type. In Petrov's book [20] a) and b) are the metrics (32.4) and (32.10) and c) is omitted.

2. If $N_1 = 0$, ($\underset{\sim}{X}_2$, $\underset{\sim}{X}_3$, $\underset{\sim}{Y}$) always generate a multiply-transitive G_3 on two-dimensional spaces, with curvature B at p. Apart from the Kantowski-Sachs case there are three possibilities
   a) $B < 0$. This is case II of [18]. It can be put in the form (3.10) with sinh θ replacing sin θ. The multiply-transitive group of Bianchi type VIII acts on two-spaces of negative curvature and the transformations $\alpha = -N_2$ with varying β and γ give a one-parameter family of groups of type III. An orthonormal basis of reciprocal group generators for one of these groups, agreeing with $\underset{\sim}{Z}_\alpha$ at p, will have $n^a{}_a = 0$ at p. The metric is (32.7) in Petrov [20].
   b) $B = 0 = A$. This can be put in the form (3.10) with θ replacing sin θ. (We can thus combine (3.10) and these last two cases together in the form

$$ds^2 = -dt^2 + X^2(t)dx^2 + Y^2(t)(d\theta^2 + f^2(\theta)d\phi^2) \qquad (3.12)$$

   where $f(\theta) = \sin\theta$, θ or $\sinh\theta$.) The multiply-transitive G_3 acts on flat two-surfaces in the case 2b, and by varying α we obtain a one-parameter family of groups of type VII_0 and a single

group of type I. This metric is (32.11) in Petrov [20].

c) $B = 0 \neq A$. Again the multiply-transitive group acts on flat two-surfaces. By varying α we obtain a one-parameter family of groups of type VII_h and a single group of type V. This metric is (32.6) in Petrov [20].

The results just obtained are identical with those in [4] and [17].

3.3. The Robertson-Walker metrics

Although this school is primarily concerned with anisotropic cosmologies, I give here (for completeness) the geometry of the Robertson-Walker models. These are the models with $s = 3$, $r = 6$.

First we may note that if the metric is spherically symmetric about every point, there is an isotropy group G_3 of spatial rotations about any point. This leaves a unique unit future-pointing timelike vector $\underset{\sim}{u}$ fixed at each point. This vector field must be geodesic and hypersurface-orthogonal, since the vectors $\dot{u}^\alpha = u^\alpha{}_{;\beta}u^\beta$ and $\omega^\alpha = 1/2\,\eta^{\alpha\beta\gamma\delta}u_\beta u_{\gamma;\delta}$ (where $\eta^{\alpha\beta\gamma\delta}$ is the totally skew pseudo-tensor of oriented volume) are orthogonal to $\underset{\sim}{u}$ and invariantly-defined, are thus fixed under the spatial rotations, and hence must vanish. The hypersurfaces to which $\underset{\sim}{u}$ is orthogonal then admit an isotropy G_3 at every point and so by Theorem 3 must be of constant curvature. This curvature may depend on time, and thus by means directly analogous to those used in deriving (3.10) we obtain the metrics

$$ds^2 = -dt^2 + R^2(t)(dr^2 + f^2(r)(d\theta^2 + \sin^2\theta d\phi^2))$$

where $f(r) = \sin r$, r or $\sinh r$ for (respectively) positive, zero and negative curvature of the three-spaces, the curvature being $k/R^2 = K$ where $k = \pm 1$ or 0.

The structure of the group G_6 can be found by direct integration of Killing's equations, or by Schmidt's method, or otherwise. It turns out, using the Schmidt method, that one can always put the commutators in the form

$$\left[\underset{\sim}{Y}_i, \underset{\sim}{Y}_j\right] = \varepsilon_{ijm}\underset{\sim}{Y}_m, \qquad \left[\underset{\sim}{Y}_i, \underset{\sim}{X}_j\right] = \varepsilon_{ijm}\underset{\sim}{X}_m$$
$$\left[\underset{\sim}{X}_i, \underset{\sim}{X}_j\right] = \varepsilon_{ijm}{}^k\underset{\sim}{Y}_m$$

When $k = 1$, there are two simply-transitive groups of Bianchi type IX. When $k = 0$ there is a group of Bianchi type I and a three-parameter family of groups of type VII_0. When $k = -1$, there are a two-parameter

family of groups of type V and a three-parameter family of groups of type VII_h [4].

The upshot of the last two sections is that the Kantowski-Sachs metric and Bianchi metrics exhaust the class of spatially-homogeneous metrics.

4. The field equations of spatially-homogeneous cosmologies

4.1. Introduction and computation

The most obvious and important fact about spatially-homogeneous cosmologies is that everything of physical importance depends only on time. Thus the field equations reduce to ordinary differential equations, and the initial conditions to the assigning of values to various constants (rather that assigning functions on a hypersurface under certain constraints). Such system of equations may be amenable to a number of well-known techniques in the qualitative theory of differential equations, in particular reduction to autonomous systems and study of Hamiltonian and Lagrangian formulations. These matters are discussed below. In this first section I aim merely to give the actual forms of the field equations. I have done the computation using Cartan's method based on differential forms. One takes a basis σ^α of differential forms and writes the metric as

$$ds^2 = g_{\alpha\beta}\,\sigma^\alpha\sigma^\beta$$

Connection one-forms $\Gamma^\alpha{}_\beta$ are then defined by

$$\begin{aligned} dg_{\alpha\beta} &= \Gamma_{\alpha\beta} + \Gamma_{\beta\alpha} \\ d\sigma^\alpha &= -\Gamma^\alpha{}_\beta \wedge \sigma^\beta \end{aligned} \tag{4.1}$$

indices being raised and lowered with $g_{\alpha\beta}$. The curvature two-form is then defined by

$$\Omega^\alpha{}_\beta = d\Gamma^\alpha{}_\beta + \Gamma^\alpha{}_\gamma \wedge \Gamma^\gamma{}_\beta = \frac{1}{2} R^\alpha{}_{\beta\gamma\delta}\,\sigma^\gamma \wedge \sigma^\delta \tag{4.2}$$

If coordinates have not been given, then to guarantee their existance one must use the condition

$$d^2\sigma^\alpha = 0 \iff R^\alpha{}_{[\beta\gamma\delta]} = 0$$

which is equivalent to the Jacobi identity (2.4) for the dual basis to σ^α. (The Bianchi identities, incidentally, are $d^2\Gamma^\alpha{}_\beta = 0$.) For

Bianchi models, which includes all the models of Chapter 3 and the Kantowski-Sachs metric (3.10), using the basis (ω^a, dt) with ω^a as in Table III, we find

$$\begin{aligned} &\Gamma^4_{\ 4} = 0, \quad \Gamma^a_{\ 4} = \theta^a_{\ b}\,\omega^b, \quad \Gamma^4_{\ a} = \theta_{ab}\,\omega^b \\ &\Gamma^a_{\ b} = \theta^a_{\ b}\,\omega^4 + \tfrac{1}{2}\,(C^a_{bc} - g_{bd}g^{ae}C^d_{ec} - g_{cd}g^{ae}C^d_{eb})\,\omega^c \end{aligned} \tag{4.3}$$

and the field equations become

$$R_{\alpha\beta} = T_{\alpha\beta} - \tfrac{1}{2}\,Tg_{\alpha\beta} + \Lambda g_{\alpha\beta} \tag{4.4}$$

where $R_{44} = -\dot{\theta} - \theta_{ab}\theta^{ab}$

$$\begin{aligned} R_{4a} &= \theta^b_{\ c}C^c_{\ ba} + \theta^c_{\ a}C^b_{\ cb} \\ R_{ab} &= \dot{\theta}_{ab} + \theta\theta_{ab} - 2\theta_{ac}\theta^c_{\ b} + {}^3R_{ab} \end{aligned} \tag{4.5}$$

Here $\omega^4 = dt$, $\dot{}$ denotes $\frac{\partial}{\partial t}$, $\theta_{ab} = (\mathcal{L}_{\underset{\sim}{n}}g)_{ab} = \frac{\partial}{\partial t}(g_{ab})$, $\theta = \theta^a_{\ a}$, $T = T^a_{\ a}$ and ${}^3R_{ab}$ is the curvature of the three-dimensional space section which is

$$\begin{aligned} {}^3R_{ab} = {} & \tfrac{1}{2}\,C^e_{\ ef}(C^g_{\ hb}g^{fh}g_{ag} + g_{bg}C^g_{\ ha}g^{fh}) + \\ & + \tfrac{1}{4}\,C^e_{\ fg}C^h_{\ ij}g_{ae}g_{bh}g^{gi}g^{fj} - \tfrac{1}{2}\,C^c_{\ da}(C^d_{\ cb} + g_{ce}g^{fd}C^e_{\ fb}). \end{aligned} \tag{4.6}$$

Note that θ_{ab} and $\frac{\partial}{\partial t}(\theta_{ab}) = (\mathcal{L}_{\underset{\sim}{n}}\theta)_{ab}$ are tensors. The formulae actually look somewhat simpler if the values for $C^a_{\ bc}$ are substituted in from Table I.

The second form I shall give for Bianchi models uses the basis (3.3) - (3.6) and is

$$\begin{aligned} &\Gamma^4_{\ 4} = 0, \quad \Gamma^a_{\ 4} = \Gamma^4_{\ a} = \theta_{ab}\underset{\sim}{e}^b \\ &\Gamma^a_{\ b} = -\,\varepsilon_{abd}\Omega^d\,\underset{\sim}{e}^4 + \left\{\tfrac{1}{2}\,(\varepsilon^a_{\ bc}n^c_{\ d} + \varepsilon^a_{\ dc}n^c_{\ d} - \varepsilon_{bdc}n^{ca}) + \right. \\ &\qquad \left. - a^a g_{bd} - a_b g^a_{\ d}\right\}\underset{\sim}{e}^d \end{aligned}$$

yielding

$$\begin{aligned} R_{44} &= -\,\theta - \theta^{ab}\theta_{ab} \\ R_{4d} &= \varepsilon_{dbc}n^{cf}\theta^b_{\ f} - 3\,\sigma_d^{\ b}a_b \end{aligned}$$

$$R_{cd} = \dot{\theta}_{cd} + \theta\theta_{cd} + 2\theta^{f}{}_{(c}\,\varepsilon_{d)ef} + 2\,\varepsilon_{fe(c}n_{d)}{}^{e}a^{f} + \\ + 2n_{(c}{}^{f}n_{d)f} - n^{e}{}_{e}n_{cd} - \delta_{cd}(2a_{e}a^{e} + n^{ef}n_{ef} - \tfrac{1}{2}(n^{e}{}_{e})^{2}). \tag{4.7}$$

In this case we have to consider the Jacobi identities, since the $\underset{\sim}{e}^{a}$ have not been given in terms of coordinates. If we use $J\{\alpha\beta\}$ to denote the equation $\sigma^{\alpha}\wedge d\sigma^{\beta} = 0$, we find we have, as non-trivial Jacobi identities,

$$\begin{aligned} &J\{a\ 4\} && : n^{db}a_{b} = 0 \\ &J\{a\ b\} + J\{b\ a\} && : \dot{n}^{ab} - 2n^{c(a}\varepsilon^{b)}{}_{cd}\Omega^{d} - 2n_{c}^{(a}\theta^{b)c} + n^{ab}\theta = 0 \\ &J\{a\ b\} - J\{b\ a\} && : \dot{a}_{b} + \theta_{b}{}^{c}a_{c} - \varepsilon_{bcd}a^{c}\Omega^{d} = 0 \end{aligned} \tag{4.8}$$

For the Kantowski-Sachs metric we calculate in the form (3.12) using the basis $(X\,dx,\ Y\,d\theta,\ Yf(\theta)\,d\phi,\ dt)$ and obtaining

$$\underset{\sim}{\Gamma}{}^{4}_{1} = \underset{\sim}{\Gamma}{}^{1}_{4} = \frac{\dot{X}}{X}\,\omega^{1}, \quad \underset{\sim}{\Gamma}{}^{4}_{2} = \underset{\sim}{\Gamma}{}^{2}_{4} = \frac{\dot{Y}}{Y}\,\underset{\sim}{\omega}{}^{2}$$

$$\underset{\sim}{\Gamma}{}^{4}_{3} = \underset{\sim}{\Gamma}{}^{3}_{4} = \frac{\dot{Y}}{Y}\,\omega^{3}, \text{ and } \underset{\sim}{\Gamma}{}^{2}_{3} = -\underset{\sim}{\Gamma}{}^{3}_{2} = \frac{-1}{Yf}\frac{\partial f}{\partial\theta}\,\underset{\sim}{\omega}{}^{3}$$

as the only non-zero $\Gamma^{\alpha}_{\beta\delta}$. Thus we obtain

$$\begin{aligned} R_{44} &= -\frac{\ddot{X}}{X} - \frac{2\ddot{Y}}{Y} \\ R_{11} &= \frac{\ddot{X}}{X} + 2\,\frac{\dot{X}\dot{Y}}{XY} \\ R_{22} &= R_{33} = \frac{\ddot{Y}}{Y} + \frac{\dot{X}\dot{Y}}{XY} + \frac{\dot{Y}^{2}}{Y^{2}} + \frac{k}{Y^{2}} \end{aligned} \tag{4.9}$$

as the only non-zero Ricci curvature components, where $k = 1$, 0 or -1 according as $f = \sin\theta$, θ or $\sinh\theta$.

If we write $T_{\alpha\beta} = \rho\, n_{\alpha}n_{\beta} + 2n_{(\alpha}q_{\beta)} + p(g_{\alpha\beta} + n_{\alpha}n_{\beta}) + \pi_{\alpha\beta}$ where $q_{\alpha}n^{\alpha} = 0$, $\pi_{\alpha\beta}n^{\beta} = 0$, $\pi^{\alpha}{}_{\alpha} = 0$, $\pi_{\alpha\beta} = \pi_{(\alpha\beta)}$ then

$$\begin{aligned} R_{44} &= \frac{\rho + 3p}{2} - \Lambda \\ R_{4d} &= -q_{d} \\ R_{ab} &= \frac{(\rho - p)}{2}\,g_{ab} + \pi_{ab} \end{aligned} \tag{4.10}$$

4.2. Degrees of freedom

Having chosen the canonical forms of Tables I and III, the 12 variables g_{ab} and θ_{ab} are related by four constraints, namely $G_{4\alpha} + \Lambda g_{4\alpha} = T_{4\alpha}$, $G_{\alpha\beta}$ being the Einstein tensor. There is still one degree of freedom in the choice of the initial surface, and a further (9 - d) degrees of freedom in choice of the ω^α. If there are M matter variables, one has in general 12 - 4 - 1 - (9 - d) + M = M + d - 2 true degrees of freedom. However this result is modified in three cases where the constraints are not necessarily independent namely where the three formulae for R_{4b} do not involve linearly independent combinations of the θ_{ab}. Computing from (4.7) using (3.6) we find

$$R_{41} = -2a(\sigma_{22} + \sigma_{33}) + \sigma_{23}(n_2 - n_3)$$

$$R_{42} = 3a\sigma_{12} + \sigma_{13}(n_3 - n_1)$$

$$R_{43} = 3a\sigma_{13} + \sigma_{12}(n_1 - n_2)$$

where $\sigma_{ab} = \theta_{ab} - \frac{1}{3}\theta g_{ab}$ (so $\sigma_{11} + \sigma_{22} + \sigma_{33} = 0$). The cases $n_2 = n_3$, $n_1 = n_2$ and $n_1 = n_3$ only arise if the full group is a G_4, as one can readily prove using (4.8). R_{42} and R_{43} give linearly dependent conditions if $(n_3 - n_1)(n_1 - n_2) + q_a^2 = 0$ i.e. if $a \neq 0$ and $h = -1/9$, or if $a = 0$ and $n_2 n_3 = 0$. The first occurs in type $VI_{-1/9}$ where R_{41} is still independent, and the second occurs in types II (where R_{41} and R_{42} are independent) and I, where none are independent.

If there are matter variables q_a as in (4.10) linesr dependence of R_{a4} imposes conditions on these and the total number of free variables is unaltered. But when the matter variables q_a are, a priori, absent, the number of constraints is decreased to 3 (types $VI_{-1/9}$ and II) or 1 (type I) and the number of free variables correspondingly increased. To show how this works out, Table IV gives the number of degrees of freedom for vacuum, and for perfect fluid with energy-momentum tensor

$$T_{\alpha\beta} = \mu u_\alpha u_\beta + p(g_{\alpha\beta} + u_\alpha u_\beta) \qquad (4.11)$$

where $\underset{\sim}{u}$, the four-velocity, need not be equal to $\underset{\sim}{n}$. In the latter case we assume p is a function of ρ and there are thus four matter variables ρ, u_a (u_4 being determined by the fact that $\underset{\sim}{u}$ is a unit vector).

The results given here are due to Siklos [7]: Λ has been treated as known, rather than a free parameter.

Table IV. Degrees of freedom of Bianchi models

Type	I	II	VI_o	VII_o	VIII	IX	V	IV	VI_h	VII_h	$VI_{-1/9}$
q	1	2	3	3	4	4	1	3	3(4)	3(4)	4
s	2	5	7	7	8	8	5	7	7(8)	7(8)	7

q=degrees of freedom of vacuum model. s degrees of freedom of fluid model. The ambiguity in the columns for VI and VII arises from that of Table II. The column for type VI_h includes type III but not h = $= -1/9$.

A similar analysis can be made for the rotationally symmetric cases of Section 3.2. For the metric (3.12) there are four variables X, $\dot{X}$, Y, $\dot{Y}$; Y is fixed except when $f(\theta) = \theta$, but X can be scaled. The field equations (4.9) imply one constraint in vacuum models, and four for fluid models. The origin of time is a free parameter.

For the remaining cases, there are still four variables but no basis change is permitted in the Bianchi VIII and IX cases, and only a rescaling of $\underset{\sim}{W}_1$ in Bianchi II and of $\underset{\sim}{W}_2$ and $\underset{\sim}{W}_3$ in Bianchi V. Here we are refering to the unique group of these types and ignoring the one parameter families of other groups. There is one constraint in the vacuum case and four in the fluid case, except for Bianchi V which has two in the vacuum case. For these cases we arrive at Table V.

Table V. Degrees of freedom of the models of Section 3.2

Type	Kantowski-Sachs	1a	1b	1c	2a	2b	2c
q	1	1	2	2	0	1	0
s	2	2	3	3	1	2	2

q and s as in Table IV. There are in fact two solutions in case 2a vacuum, but no continuous parameter: one of these solutions is flat space in unusual coordinates.

4.3. Diagonalizability

The simplest metrics are those where Θ_{ab}, $T_{\alpha\beta}$ and n_{ab} are simultaneously diagonal, and $\Omega_b = 0$, in (4.7). The conditions under which such a form is enforced by the governing equations have been investigated [21, 22]. The proof works by use of the R_{4a} equations, the Jacobi identities, the Bianchi identities and lastly the R_{ab} equations ($b \neq c$). This route breaks down in types I, II, and VI ($h = -1/9$) because of the linear dependence of the R_{4a} equations.

Lemma 4: In class A, except types I and II, the condition $q_a = 0$ implies n_{ab}, π_{ab} and Θ_{ab} are diagonal and $\Omega_d = 0$.

Proof [21]: $R_{41} = 0$ implies $n_{23}(\theta_2 - \theta_3) = 0$, using a basis (3.3) - (3.5) chosen so that θ_{ab} is diagonal. So either $n_{23} = 0$ or $\theta_2 = \theta_3$. In the latter case the basis is not yet fixed and we can impose $n_{23} = 0$ as a further condition on it.

Cyclic permutation implies n_{ab} can be taken diagonal. Then $J\{1\ 2\} + J\{2\ 1\}$ yields $(n_{22} - n_{11})\Omega_3 = 0$. Thus if $n_{22} \neq n_{11}$, $\Omega_3 = 0$. If $n_{22} = n_{11}$ then $J\{1\ 1\}$ and $J\{2\ 2\}$ give $\Theta_{11} = \Theta_{22}$ and we can impose $\Omega_3 = 0$ as the restriction on the still indeterminate basis. Cyclic permutation of this argument yields $\Omega_b = 0$.

Finally R_{cd} ($c \neq d$) can be evaluated. It is zero and thus $\pi_{cd} = 0$.

Lemma 5: In Bianchi types I and II any three of the conditions i) n_{ab} diagonal, ii) Θ_{ab} diagonal, iii) π_{ab} diagonal and iv) $\Omega_d = 0$ imply the fourth.

Proof [21]: As in Lemma 4 n_{ab} and Θ_{ab} can be taken to be simultaneously diagonal. Then $\Omega_d = 0$ if and only if $R_{cd} = 0$ ($c \neq d$) which in turn is true if and only if $\pi_{cd} = 0$, using residual freedom of rotation of the basis (3.3) if necessary.

Lemma 6: In class B, except type VI ($h = -1/9$), $q_b = 0$ implies a_b is an eigenvector of Θ_{ab} and π_{ab}, and is parallel to Ω_d.

Proof [21]: Use the basis (3.3) - (3.6). Then $R_{42} = R_{43} = 0$ gives $\sigma_{12} = \sigma_{13} = 0$ and $R^{\alpha\beta}{}_{;\beta} = 0$ gives $\pi_{12} = \pi_{13} = 0$. Then $J\{1\ 2\} - J\{2\ 1\}$ and $J\{1\ 3\} - J\{3\ 1\}$ give $\Omega_2 = \Omega_3 = 0$.

Lemma 7: In class B, except type VI ($h = -1/q$), if $T_{\alpha\beta}$, Θ_{ab} are simultaneously diagonal and $\Omega_d = 0$, the space-time is either rotationally symmetric, or $n^a{}_a = 0$, or the matter content is physically unrealistic.

Proof [22]: Use the basis in which $a_b = (a, 0, 0)$ and Θ_{ab} is diagonal (this exists by Lemma 6). Then $J\{a\ 4\}$ gives $n_{11} = n_{12} = n_{13} = 0$

$$R_{23} = 0 \quad \text{gives} \quad (n_{22} - n_{33})a + (n_{22} + n_{33})\, n_{23} = 0$$

The time derivative of this gives

$$(\theta_2 - \theta_3)\ (n_{22} + n_{33})a_2 + (n_{22} - n_{33})n_{23} = 0$$

so either i) $(n_{22} - n_{33})^2 = (n_{22} + n_{33})^2$, whence $n_{22}n_{33} = 0$ so (without loss of generality) take $n_{22} = 0$ and then $R_{23} = 0$ implies either $n_{33} = n^a{}_a = 0$ or $a = n_{23}$. In the latter case $R_{14} = 0$ gives $\theta_{11} = \theta_{22}$ and $R_{11} = R_{22}$ and hence $\pi_{11} = \pi_{22}$. This model is rotationally symmetric or ii) $\theta_2 = \theta_3$. Then $R_{14} = 0$ gives $\theta_1 = \theta_2 = \theta_3$. If t is defined by $\theta_1 = 1/t$, all the n_{ab} and a_b are proportional to $1/t$ and thus π_{ab} is proportional to t^{-2}, but this violates the "dominant energy" condition at large t unless $\pi_{ab} = 0$, which gives ${}^3R_{ab} = \frac{{}^*R}{3} g_{ab}$ implying the three-spaces have constant curvature and thus the models are Robertson-Walker.

It may be noted that the result of Lemma 7 is true for type VI ($h = -1/9$) but the proof [22] is more tedious.

4.4. Restrictions on the matter content

The ideas of Section 4.2 and 4.3 can be re-worked to show that diagonal metrics (where n_{ab} and θ_{ab} are simultaneously diagonal) always have $q_b = 0$ in Class A. In type I, $q_b = 0$ always and in type II q_b has only one free component. In the models of Section 3.2, only the type V case, 2c, allows $q_b \neq 0$.

King and Ellis [12] have extensively investigated the case of a perfect fluid with $\underset{\sim}{u} \neq \underset{\sim}{n}$. In this case they wrote

$$\begin{aligned} \underset{\sim}{u} &= \underset{\sim}{n} \cosh\Phi + \underset{\sim}{c} \sinh\Phi \\ \underset{\sim}{n} &= \underset{\sim}{u} \cosh\Phi + \hat{\underset{\sim}{c}} \sinh\Phi \end{aligned} \qquad (4.12)$$

where $\underset{\sim}{c}$, $\hat{\underset{\sim}{c}}$ are spacelike unit vectors orthogonal to $\underset{\sim}{n}$ and $\underset{\sim}{u}$ respectively. The homogeneity implies that Φ is a function of t alone so we can write $\Phi_{,a} = -\dot{\Phi} n_a$.

Taking the energy-momentum tensor to be (4.11) with $p = p(\mu)$ we find

$$\mu_{;\alpha}u^\alpha + (\mu + p)u^\alpha{}_{;\alpha} = 0$$

$$(\mu + p)u^\alpha{}_{;\beta}u^\beta + (g^{\alpha\beta} + u^\alpha u^\beta)p_{,\beta} = 0$$

So, defining

$$w = \exp\int \frac{d\mu}{\mu + p}\ , \quad r = \exp\int \frac{dp}{\mu + p}$$

the integrals being taken along the worldlines with initial values in a surface of homogeneity, we have

$$\cosh\phi\ (\log w)^{\bullet} + u^{\alpha}{}_{;\alpha} = 0 \tag{4.13a}$$

$$\sinh\phi\ (\log r)^{\bullet} c_{\alpha} + u_{\alpha;\beta} u^{\beta} = 0 \tag{4.13b}$$

where $^{\bullet}$ denotes d/dt as usual. Thus

$$u^{\alpha}{}_{;\beta} u^{\beta} = \tanh\phi\ \frac{dp}{d\mu}(u^{\beta}{}_{;\beta}) c^{\alpha} \tag{4.14}$$

this shows that in a tilted universe $u^{\alpha}{}_{;\alpha} \neq 0$ (expansion of the fluid) implies either $u^{\alpha}{}_{;\beta} u^{\beta} \neq 0$ (acceleration) or $dp/d\mu = 0$. Substituting the formula for $\underset{\sim}{u}$ into the equation (4.13) yields

$$(\log (r \sinh\phi))^{\bullet} + c^{\alpha} \Theta_{\alpha\beta} c^{\beta} = 0$$

so that tilted models stay tilted.

The vorticity of a fluid is given by the vector

$$\omega^{\alpha} = \frac{1}{2}\ \eta^{\alpha\beta\gamma\delta} u_{\beta} u_{\gamma;\delta}$$

(see e.g. [23, 24] for an introduction to relativistic fluid dynamics in cosmology). Since $n_{[\alpha;\beta]} = 0$, and $\phi_{,\beta}$, n_{β} and u_{β} are coplanar, this gives

$$\omega^{\alpha} = \frac{1}{2}\ \sinh\phi\ \eta^{\alpha\beta\gamma\delta} u_{\beta} c_{\gamma;\delta}$$

Because $\omega^{\alpha} u_{\alpha} = 0$ we will have

$$\omega^{\alpha} n_{\alpha} \cosh\phi + \omega^{\alpha} c_{\alpha} \sinh\phi = 0$$

so ω^{α} is given (when $\phi \neq 0$) uniquely by the components ω^{a}, which are in turn given by the components of $\varepsilon^{abe} c_{[a;b]}$ (where Latin indices refer to a basis in the orbits). These can be written in terms of

$$(n^{ed} c_{d} + \varepsilon^{edf} c_{d} a_{f}) \tag{4.15}$$

From this expression, we can deduce the following:

a) Bianchi type II models have zero vorticity, and velocity (from (4.5)) in the plane of $\underset{\sim}{u}$ and the eigenvector of n_{ab} with non-zero eigenvalue.

b) In Class B models, if a_{f} and c_{d} are parallel, $\omega^{\alpha} = 0$.

c) In Class B, if a_{f} and c_{d} are not parallel, $\omega^{\alpha} = 0$ can only occur in Bianchi type III.

d) In Bianchi types VIII and IX, tilt implies vorticity.

e) In Bianchi types VII and IX, $\underset{\sim}{\varsigma}$ and $\underset{\sim}{\omega}$ cannot be orthogonal, and hence $\omega^{\alpha} n_{\alpha} \neq 0$.

f) In Bianchi type V, $\underset{\sim}{\varsigma}$ and $\underset{\sim}{\omega}$ are orthogonal, while $\underset{\sim}{\varsigma}$ and $\underset{\sim}{\omega}$ can be non-zero and orthogonal in Bianchi type IV (this requires two eigenvectors of $n_{\alpha\beta}$ with zero eigenvalue). In this cases $\underset{\sim}{\omega}$ is orthogonal to $\underset{\sim}{n}$, $\underset{\sim}{a}$, and $\underset{\sim}{\varsigma}$.

g) $\underset{\sim}{\varsigma}$ and $\underset{\sim}{\omega}$ are parallel and non-zero only if $a_b c^b = 0$. Further investigation shows this can only occur in Class A solutions with $\underset{\sim}{\varsigma}$ an eigenvector of n_{ab}.

h) The only tilted models with a G_4 on spacelike surfaces are the Bianchi V models.

All these results are given by King and Ellis [12].

Another form of energy-momentum that may be considered is a Maxwell electromagnetic field. Maxwell's equations read (in terms of the usual electric and magnetic fields as seen by an observer with velocity n^a, and using the basis (3.3) - (3.6))

$$2E_b a^b = j^{\alpha} n_{\alpha}$$

$$a_b H^b = 0$$

$$(E^c)^{\cdot} = - j^c + \Theta^c{}_b E^b - \Theta E^c + n^{cd} H_d + \varepsilon^{ceb} a_e H_b$$

$$(H^d)^{\cdot} = \Theta^c{}_d H^d - \Theta H^c - n^{cd} E_d - \varepsilon^{ceb} a_e E_b .$$

We see that for a pure magnetic field,

$$a_b H^b = 0$$

$$n^{cd} H_d + \varepsilon^{ceb} a_e H_b = 0$$

which shows that no field is possible except in types I, II (2 free components), and III, VI_o and VII_o (1 free component). This result is due to Hughston and Jacobs [25]; see also [26]. Similar considerations apply to massive vector-meson fields [25].

The other forms of anisotropic stress that can arise derive from either a kinetic theory treatment of some particle distribution, or from an approximative treatment of gravitational wave fluxes (at least, those are the cases considered in the literature). I therefore leave their discussion, apart from the broad restrictions above, to the other lectures.

The Kantowski-Sachs metric cannot be tilted and only one component of H_b can be non-zero.

4.5. Lagrangian and Hamiltonian forms

The Einstein tensor can be derived by varying $g_{\alpha\beta}$ in the Lagrangian action

$$I = \int R \sqrt{-g}\, d^4x \tag{4.16}$$

where $\sqrt{g}$ is the determinant of $g_{\alpha\beta}$. If the energy-momentum tensor can be derived from a Lagrangian form also, then one will have a completely Lagrangian description.

Misner introduced idea of taking the Lagrangian (4.16) and integrating over the spatial variables in a spatially-homogeneous model (taking the spatial volume finite) to obtain

$$I = \int R \sqrt{-g}\, dt \tag{4.17}$$

where R and g now depend on t alone (i.e. here $g_{\alpha\beta}$ is refered to a group invariant basis). This was used for Bianchi types I [27] and IX [28], and the corresponding Hamiltonian formulation was introduced, for type IX [29], and used extensively, see e.g. [34]. Several authors attempted to generalize the technique to other Bianchi types [30, 5], but it turns out there is a snag. I myself, and independently Estabrook and Wahlquist, found that the generalization of (4.17) to types of Class B did not in general seem to work. The reason for this was suggested by Hawking [31] and given in full by myself and Taub [32]. Informed of our work by Ehlers, Ryan checked the Hamiltonian form and found the same difficulty, but owing to an error ascribed the result to the wrong cause [33]. It was already known, also, that there were certain special cases of Class B, namely those in which, in terms of (3.6), $n^a{}_a = 0$, where the Lagrangian treatment can be applied, and Taub and I [32] had tried to analyze the ways of amending the Lagrangian to make it work. Unfortunately, in doing so, I had miscalculated and given an incorrect statement. This amusing history of successive errors ends (at least for now!) with the work of Sneddon [35], who corrected all the errors mentioned and gave a clear proof showing why the method can be made to work for the $n^a{}_a = 0$ cases.

After all that historical information, I must now explain the problem. Rather than repeat all the tedious algebra, I shall use a simpler example. Consider a Lagrangian action

$$I = \int_{x_1}^{x_2} f(x,y,y')dx$$

where $y' = \partial y/\partial x$, and the variation $\delta I = 0$ for an arbitrary varia-

tion δy in y.

$$\delta I = \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y} \delta y + \frac{\partial f}{\partial y'} \delta y' \right) dx = \left[\frac{\partial f}{\partial y'} \delta y \right]_{x_1}^{x_2} + \int_{x_1}^{x_2} \delta y \left(\frac{\partial f}{\partial y} - \frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right) \right)$$

where the first term denotes a definite integral. To obtain the usual Euler-Lagrange equations one must remove these boundary terms. Usually this is done by setting $\delta y = 0$ at x_1 and x_2, but the alternative is to impose the "natural boundary conditions" $\partial f / \partial y' = 0$ at x_1 and x_2.

The analogy is that in using the Lagrangian (4.16) one varies over a four-dimensional region, with the variation $\delta g_{\alpha\beta}$ set to zero on a three-dimensional boundary. However, this cannot be done in a spatially-homogeneous way, because if $\delta g_{\alpha\beta} = 0$ on the boundary it is zero for every t and hence zero everywhere. Therefore one can only use the method, with spatially-homogeneous variations, when the natural boundary conditions are satisfied. These turn out to be that the group is Class A.

When the method fails, (4.17) will give

$$\delta I = \int (E^{\alpha\beta} + V^{\alpha\beta}) \, \delta g_{\alpha\beta} \, dt$$

where $E^{\alpha\beta} = 0$ are the Einstein equations, and $V^{\alpha\beta}$ are the unwanted additional terms. By taking a suitable basis of $\delta g_{\alpha\beta}$, considered as a vector space, one can arrange that only one component of $V^{\alpha\beta}$ is non-zero (it was on this point that [32] went wrong). Thus there is one incorrect equation, which can be corrected by imposing the true equation as a (non-holonomic) constraint. The remaining issue is to discover all cases where this constraint becomes holonomic. The $n^a_a = 0$ cases do belong to this category. Dr. M. Francaviglia has suggested these may be exactly the cases where it is possible to find a compact global topology for the space sections without losing global homogeneity.

I shall not give the most general forms of Lagrangians (for which see [34]), but treat only the case of Class A models in vacuum. Then if we write

$$g_{ab}(t) = e^{2\lambda} (e^{-2\beta})_{ab} \qquad (4.18)$$

where g_{ab} is diagonal (by Section 4.3), and β_{ab} is a diagonal trace-free matrix, we can express the result in terms of λ, β_+ and β_-

where

$$\beta_{ab} = \mathrm{diag}\,(\beta_+,\ \sqrt{3}\,\beta_- - \frac{\beta_+}{2},\ -\sqrt{3}\,\beta_- - \frac{\beta_+}{2}).$$

We find a Lagrangian

$$L = ({}^3R + 6\dot{\lambda}^2 - \frac{3}{2}(\dot{\beta}_+^2 + \dot{\beta}_-^2))e^{3\lambda}$$

where the scalar curvature of the three spaces is

$${}^3R = -\frac{e^{-2\lambda}}{2}(N_1^2\, e^{4\beta_+} + e^{-2\beta_+}(N_2 e^{\sqrt{3}\beta_-} - N_3 e^{-\sqrt{3}\beta_-})^2 -$$

$$- 2N_1 e^{\beta_+}(N_2 e^{\sqrt{3}\beta_-} + N_3 e^{-\sqrt{3}\beta_-})) + \frac{1}{2}N_1N_2N_3(1 + N_1N_2N_3)$$

with N_1, N_2, N_3 as in Table I. Using λ as a new time-variable this has the form of the Lagrangian of a particle in a two-dimensional potential which has exponential behaviour and is contracting as λ increases. Some sketches for the non-trivial potentials of types II, VI_0, VII_0, VIII and IX are given as Figs. 1 - 5.

4.6. Autonomous systems of equations

A system of ordinary differential equations is said to be autonomous if it takes the form

$$\frac{d\underset{\sim}{x}}{d\lambda} = \underset{\sim}{f}(\underset{\sim}{x})$$

where $\underset{\sim}{x}$ is some n-tuple of variables and $\underset{\sim}{f}$ some n-tuple in $\underset{\sim}{x}$, independent of λ. The evolution then follows a curve in $\underset{\sim}{x}$-space whose principal features can be found by a study of the geometry of the vector field $\underset{\sim}{f}$, in particular of its zeros and singularities.

A plane autonomous system is one in which $\underset{\sim}{x}$-space is a plane. Well-known results due to Poincare and Bendixson make this case very easy to treat. Tables IV and V reveal the following possible cases, which may be treated.

Type I with fluid [36][37] (including $\Lambda \neq 0$)
Type II vacuum [36]
Kantowski-Sachs with fluid [17][38]
Type II, rotationally symmetric, with fluid [36]
Type III, rotationally symmetric, with fluid [36]
Type V, rotationally symmetric, with fluid [39][40]

Most of these have been treated by Collins (see references) and a few by others. The type I models have also been treated with magnetic

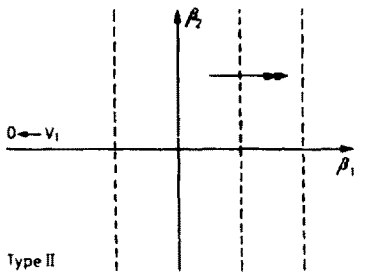

Fig. 1

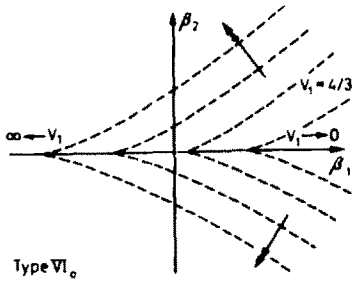

Fig. 2

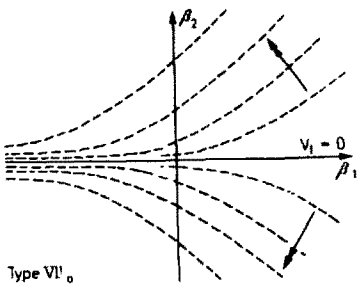

Fig. 3

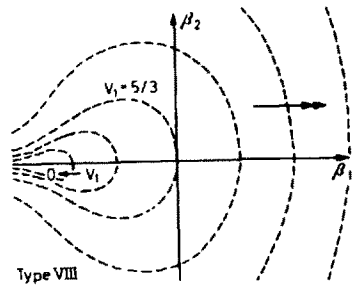

Fig. 4

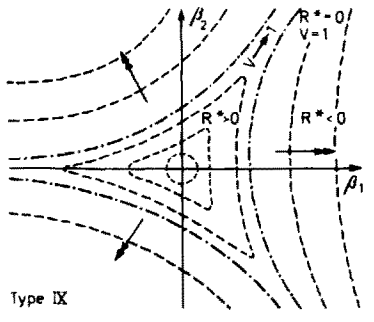

Fig. 5

Figs. 1 - 5. Sketches for the non-trivial potentials of types II, VI_o, VII_o, VIII, IX are given. The dotted lines are contours of ${}^3R$. The double-headed arrows are directions of exponential increase. ${}^3R > 0$ except in the central region of Fig. 5.

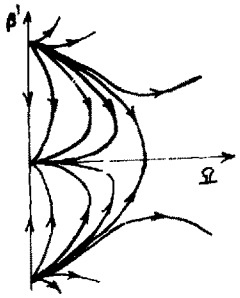

Fig. 6. Evolution for Bianchi I fluid models with $\Lambda = 0$ and $1 \leqslant \gamma \leqslant 2$.

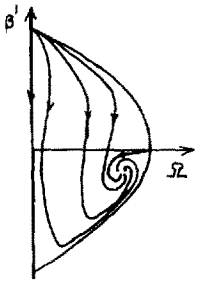

Fig. 7. Evolution for rotationally-symmetric type II model with fluid $1 \leqslant \gamma \leqslant 2$.

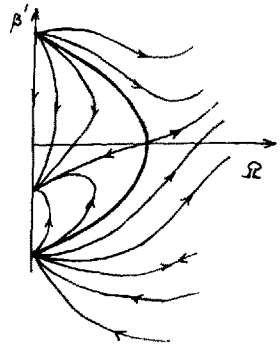

Fig. 8. Evolution of Bianchi type VI_h model ($n^a{}_a = 0$) with fluid, $4 \leqslant (1 - 3h)(3\gamma - 2)$; the outer region shows the Kantowski-Sachs evolution which joins naturally on to type III ($n^a{}_a = 0$) which is type VI_h ($h = -1$).

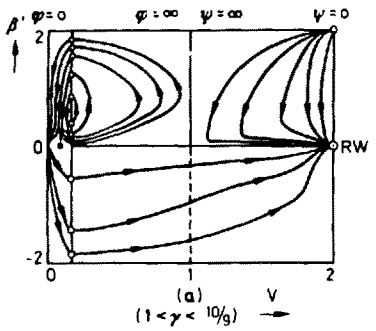

Fig. 9. Evolution of type V rotationally-symmetric tilted fluid model with $1 < \gamma < 10/9$.

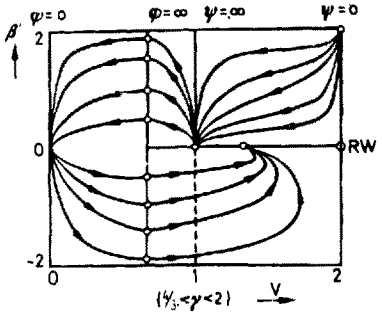

Fig. 10. As for Fig. 9 with $4/3 < \gamma < 2$.

field [41], with rotational symmetry and free neutrinos [42], and with a fluid in which $p = p(\mu, u^a{}_{;a})$, i.e. with bulk viscosity [43]. The further important group of cases are those with $n^a{}_a = 0$ where the extra cases (types VI_0 and VI_h) turn out, with perfect fluids, to have 2 free variables. These were studied by Collins also [36].

To turn the equations into plane autonomous systems, it is best to use the λ of Section 4.5, a density variable $\Omega = 3\mu/\theta^2$ describing the dynamical importance of matter, and a variable defined by

$$\beta' = \frac{d\beta}{d\lambda} = \sqrt{\beta_+'^2 + \beta_-'^2} = 2\sqrt{3}\,\sigma/\theta$$

where σ is the shear. These parameters have been renamed, and changed compared with the original papers so that λ proceeds in the direction of cosmological expansion and Ω is the same as the density parameter now generally used in FRW models. λ will be doubl-valued if the universe recollapses. A model is FRW if $\beta = 0$ throughout its evolution. It is matter-dominated when $\Omega = 1$ and matter is negligible if $\Omega = 0$. Figs. 6 - 10 show some examples from the various cases. In these the fluid obeys

$$p = (\gamma - 1)\mu \qquad (4.19)$$

The type VI_0 fluid $n^a{}_a = 0$ case and type VI_h fluid with $n^a{}_a = 0$, $4 > (1 - 3h)(3\gamma - 2)$ with $\gamma < 2$, have similarities with Fig. 7. All types show structural instability with respect to γ at $\gamma = 2$ [36]. In the type V rotationally symmetric titled fluid model there are further instabilities at $\gamma = 10/9$, $6/5$ and $4/3$ [39]. To illustrate this I give two of Collins' diagrams as Figs. 9 and 10. In these figures, the filled dotes are endpoints of evolution and the horizontal axis V represents the "angle" used in (4.12), by $V = 2/(1 + \tanh^2\phi)$. Note that $V < 1$ corresponds to a stationary region in which n^a, the hypersurface normal, is spacelike.

The system which cannot be reduced to the plane autonomous case have been studied, at least for some cases, by Bogoyavlenskii and Novikov [44]. Their works is too detailed to be usefully summarized here (though it will be referred to later).

4.7. Exact solutions

A large number of exact solutions are known. The table given in [96] is now out-dated, and a full account of the solutions will appear in a forthcoming book by Stephani, Kramer, Herlt and myself. The

work for this book is not completed at the time of writing, and so I give here a table that may prove incomplete. I do not think the full details of the metrics can be appropriately given here, so I give only references to the original literature. I shall ignore FRW solutions. In the following G denotes a general solution, S a special case.

Solutions with G_4

Type I (VII_0): vacuum: Kasner [58], Taub [59], Ehlers and Kundt [60] with $\Lambda \neq 0$, G.
dust: Saunders [61] with $\Lambda \neq 0$, Heckmann-Schucking [62], Robinson [63], Doroshkievich [64], G.
"radiation": Doroshkevich [64], Shikin [65], Thorne [66], Stewart [67], Kompaneets and Chernov [19], G.
other fluids: [64], [66], [67], G.
magnetic field: Rosen [68], [64], [67], [65], [66], G.
magnetic field plus fluid: Jacobs [69], [64], G.

Type II: vacuum: Taub [59], Newman et al. [70], Stewart [67] with $\Lambda \neq 0$, Carter [71, 72], Cahen-Defrise [73], G.
dust: Collins [36], S.
"radiation" and other fluids: Collins [36], Maartens and Nel [74], S.

Type III and Kantowski-Sachs:
vacuum: Kantowski and Sachs [16, 18], Ehlers and Kundt [60], Siklos [14], Cahen and Defrise [73], G. Can have $\Lambda \neq 0$ and a G_6.
dust: Kantowski and Sachs [16, 18], with $\Lambda \neq 0$, Kompaneets and Chernov [19], G.
radiation: Kantowski [16]
others: Kantowski [16]
magnetic field: [64], [66], [67], with fluids

Type V: vacuum: none
dust: Farnsworth [75] with tilt, G.
fluid ($\gamma = 2$): Maartens and Nel [74], Wainwright, Ince and Marshnan [76], S.
electromagnetic field and fluid: Melvin [77]

Types VIII and IX:
vacuum: Taub [59] - NUT [70], with $\Lambda \neq 0$, Stewart [67], Carter [72], etc. G.
fluid: Collins, Glass and Wilkinson (unpublished), Maartens and Nel ($\gamma = 2$) [74], S.
electromagnetic field: Brill [78], [73], G.
field and fluid and Λ : Ozsvath [79], S.

Metrics with a G_3 on a spacelike sections

Type I: vacuum: Kasner [58] et al. (see above), G.
dust: Saunders [61] et al. (see above), Raychaudhuri [80], G.
radiation and other fluids: Jacobs [81], Hughston and Shepley [87], G.
magnetic field: Rosen [68, 82], [69]
magnetic field and fluid: [69]
magnetohydrodynamic: Ozsvath [83], S.
electromagnetic: Barnes [84], S.

Type II: vacuum: Taub [59], G.
dust and other fluids: Collins [36], [85], Maartens et al. [74], Wainwright et al. [76] ($p = \mu$ with tilt), S.
electromagnetic: Barnes [84], S.

Type IV: vacuum: a special plane wave (Siklos [14], Harvey and Tsoubelis [86])

Type V: vacuum: Joseph [124], G.
dust: Heckmann-Schucking [62], G.
other fluids: (in principle), Ellis et al. [4], Hughston et al. [87]

Type VI_0: vacuum: MacCallum [4], S.
dust: MacCallum [4], S.
other fluids: MacCallum [4], Collins [36], Dunn and Tupper [88], S.
electromagnetic field and fluid: [88], S.

Type VI_h ($h \neq -1/9$):
vacuum: special plane waves Siklos [14], MacCallum [4], Collins [36], S.
fluids: Collins [36], Wainwright et al. [76], S.
electromagnetic field: Barnes [84], S.

Type VI_h ($h = -1/9$):
vacuum: Petrov Type III: Collinson and French [89], Siklos [14], S.
$\Lambda \neq 0$: Petrov N, III and II solutions, Siklos [14], S.

Type VII_0: fluid: Demianski and Grishchuk [90], S.

Type VII_h: vacuum: special plane waves Siklos [14], Lukash [91], S.
fluid ($p = \mu$): Wainwright et al. [76], Barrow [92], S.

5. Effects of geometry on the evolution

5.1. Singularities

The singularity theorems due to Hawking, Penrose and others [45] guarantee that the spatially-homogeneous models possess singularities. In fact some timelike geodesics must be incomplete, provided that $R_{\alpha\beta} k^{\alpha} k^{\beta} > 0$ for all timelike and null vectors k and the Cauchy problem has a unique solution ([45], page 147).

From (4.7) we see that if the matter and spatial curvature are both negligible, the solution approximates a Bianchi I vacuum metric. These are a one-parameter family (cf. Table IV)

$$ds^2 = - dt^2 + t^{2p} dx^2 + t^{2q} dy^2 + t^{2r} dz^2 \qquad (5.1)$$

where $p + q + r = p^2 + q^2 + r^2 = 1$. Except in the special case where $p = 1$, $q = r = 0$ (and cyclic interchanges) this Kasner metric has a real singularity at $t = 0$. It is of the "cigar" or "line" type, because a co-moving region which is spherical at $t = t_1 > 0$ becomes infinitely long and thin as $t \longrightarrow 0$. The special case $p = 1$ apparently gives a singularity of a "pancake" type, where the spherical region becomes an infinitely thin disk, but the metric is actually just flat space with unusual coordinates. This will become a true singularity when the matter content is considered.

This simple approximation suggests that Bianchi universes have "big-bang" singularities. Using an Bianchi V model with rotational symmetry, Shepley showed that another possibility exists [46]. This is the presence of a "wimper" [13] or "intermediate singularity" or "non-scalar" [47] singularity, one at which no curvature invariant becomes singular, but at which in a frame parallelly-propagated along a geodesic hitting the singularity, the Riemann tensor components diverge. In our universes, such a possibility is associated with the surfaces of homogeneity becoming null. The (null) hypersurface-normals $\underset{\sim}{n}$ are still invariant under the group, and a Killing vector field $\underset{\sim}{Y}$ agreeing with $\underset{\sim}{n}$ at one point p in the null hypersurface will obey

$$\mathcal{L}_{\underset{\sim}{Y}} \underset{\sim}{n} = 0 = \mathcal{L}_{\underset{\sim}{n}} \underset{\sim}{Y}.$$

So $\underset{\sim}{Y}$ will be parallel to $\underset{\sim}{n}$ at all points on the null hypersurface generator λ starting at p with tangent vector $\underset{\sim}{n}$. We have

$$n^{\alpha}{}_{;\beta}\, n^{\beta} = -(\gamma + \bar{\gamma}) n^{\alpha}$$

(using Newman-Penrose notation) and thus if s is affine on λ and $\underset{\sim}{Y} = \partial/\partial v$ we find, taking $\underset{\sim}{k}$ as geodesic on λ

$$n^{\alpha} = (\gamma + \bar{\gamma})\, s\, k^{\alpha} = e^{-(\gamma + \bar{\gamma})v}\, k^{\alpha}$$

So at s = 0, the components of the Riemann tensor in a tetrad parallelly-propagated with respect to $\underset{\sim}{k}$ become divergent (they are constant in the group-invariant tetrad using $\underset{\sim}{n}$).

The possible cases were examined by Ellis and King [13], Collins [39] and Siklos [7]. The examples shown in Figs. 9 - 10 include models which cross a "wimper" type null surface at v = 1. The most recent work [7] indicates that "wimpers" only occur in a special subset of all models. The proof uses a null tetrad invariant under the group with n^{α} as a hypersurface-normal and l^{α} tangent to a family of null geodesics. The remaining vectors $\underset{\sim}{m}$, $\underset{\sim}{\bar{m}}$ of the Newman-Penrose [14] formalism are also invariant and hence any connection coefficient Φ is invariant under $\underset{\sim}{n}$, $\underset{\sim}{m}$ and $\underset{\sim}{\bar{m}}$ i.e. in Newman-Penrose notation it obeys

$$\Delta\Phi = \delta\Phi = \bar{\delta}\Phi = 0$$

This drastically simplifies the NP equations. Using the commutators of $\underset{\sim}{l}$, $\underset{\sim}{n}$, $\underset{\sim}{m}$ and $\underset{\sim}{\bar{m}}$ shows that

$$D e^{-2\eta} = -(\gamma + \bar{\gamma}) - (\varepsilon + \bar{\varepsilon}) e^{-2\eta}$$

$$0 = \bar{\alpha} + \beta - \bar{\pi}$$

$$\mu = \bar{\mu}$$

on the null hypersurface, where the hypersurface normals are parallel to $n_{\alpha} + e^{-2\eta} l_{\alpha}$ throughout the region considered.

(Note on the NP formalism: We use the signature (-, -, -, +) here, and $l^{\alpha} n_{\alpha} = 1 = - m^{\alpha} \bar{m}_{\alpha}$: all other cross-products are zero. Writing $\underset{\sim}{e}_1 = \underset{\sim}{m}$, $\underset{\sim}{e}_2 = \underset{\sim}{\bar{m}}$, $\underset{\sim}{e}_3 = \underset{\sim}{n}$, $\underset{\sim}{e}_4 = \underset{\sim}{l}$ the independent connection forms can be expressed as

$$\underset{\sim}{\Gamma}^1_3 = \mu \underset{\sim}{\omega}^1 + \lambda \underset{\sim}{\omega}^2 + \nu \underset{\sim}{\omega}^3 + \pi \underset{\sim}{\omega}^4$$

$$\underset{\sim}{\Gamma}^1_4 = - \rho \underset{\sim}{\omega}^1 - \sigma \underset{\sim}{\omega}^2 - \tau \underset{\sim}{\omega}^3 - \kappa \underset{\sim}{\omega}^4$$

$$\frac{1}{2} (\underset{\sim}{\Gamma}^4_4 + \underset{\sim}{\Gamma}^1_1) = \beta \underset{\sim}{\omega}^1 + \alpha \underset{\sim}{\omega}^2 + \gamma \underset{\sim}{\omega}^3 + \varepsilon \underset{\sim}{\omega}^4$$

$\underset{\sim}{\Gamma}^4_4$ is real and $\underset{\sim}{\Gamma}^1_1$ imaginary.)

Theorem 7: Class A whimpers are impossible if

$$\Phi_{22} = R_{\alpha\beta} n^{\alpha} n^{\beta} > 0$$

Proof: 7 Calculation shows the group in Class A if $\mu = 0$. The NP equation (n) gives a contradiction.
Siklos [7] has proved that there are only three disjoint two-parameter families of vacuum whimpers. Thus the whimpers are not the general case (which needs a four-parameter family). Two of the three cases are the Taub-NUT models and plane wave solutions. The proof is readily generalized to fluids [7]. It runs essentially as follows.

There are 24 real parts of the spin coefficients, Γ^i. The NP equations give $D\Gamma^i$ for 18 of these, and give 8 of the 10 real components ψ^i of the conformal curvature tensor; the remainder are constraints. The missing $D\Gamma^i$ are fixed either by choice of tetrad or through the commutators. The missing ψ^i are given by a Bianchi identity. There remain 14 constraints on the 24 Γ^i. 8 of the 10 remaining free values are, or can be, fixed by choice of tetrad, leading only 2 free. Detailed algebra gives all possibilities.

Thus we are left to consider the "big-bangs" as the general case. (This is easy to show for the Kantowski-Sachs models [17]). A full review of the arguments has been given by Collins and Ellis [98]. We must now consider the approach to the singularity.

First suppose that the energy-momentum has negligible effect, and let us consider just the models with a Lagrangian form. From Figs 1 - 5 we can see that the evolution can be approximated in such cases by a sequence of the following periods i) periods when ${}^3R$ is negligible and the behaviour is like (5.1), ii) periods when an exponential wall, as in Fig. 1, is important, iii) periods when a corner channel (as on the left of Fig. 3) is important. The behaviour in period ii) is given by Taub's vacuum Bianchi II solution. It gives a law for passing from one period i) to another (these are generally called Kasner epochs) with a change of parameter. The period iii) has been studied by numerous authors. The model is eventually reflected (to the right in Fig. 3) after numerous oscillations. Combining these elements, Belinskii, Khalatnikov and Lifschitz [100, 48] have given qualitative descriptions of the evolution of Bianchi VIII and IX universes, in terms of a series of Kasner epochs with changing parameter and therefore two length scales oscillating, punctuated by runs up corner channels and permutation of the axes of oscillation. The rigorous methods of Bogoyavlenskii and Novikov [44] show that this qualitative description is justified and it is also confirmed by numerical work [49]. Belinskii et al. [48] have gone on to argue that the behaviour found is in fact typical of a general class of inhomogeneous cosmologies (essentially found by letting n_{ab} and a_b be functions of

spatial position but slowly-varying). This work has come under heavy criticism from J. Barrow and F. Tipler (unpublished) and it has been suggested that in any case it represents only a second approximation to the true generic big-bang. Further investigation is required, and will undoubtedly be undertaken.

We have now to consider whether the matter really is negligible near the singularity. This will certainly not be true in the extreme case $p = \mu$, but there are other special solutions where it is not true. The first of these was noted by myself [5], and more were found by Collins. In Fig. 7 there are two such solutions, one at the focus of the evolution curves, and the one starting at $\Omega = 1$, $\beta' = 0$.

In type IX tilted models the matter terms contribute extra potentials [11, 34] which, as the singularity is approached, confine the solution to one of six congruent areas of Fig. 5 [50]. The type VII_h solutions are similar to the type VII_0 case [51], because although there is no Lagrangian form, the field equations actually differ only by the addition of a simple term negligible near the singularity.

5.2. Isotropization

It is very difficult to define the requirements on a model in terms of isotropy. What one really requires is just that no conflict with present-day observation arises. Beyond this, judgments are aesthetic. One criterion that gives an interesting aesthetic argument is to consider the evolution of those models which expand indefinitely, and then demand that they approach FRW behaviour (i.e. $\sigma/\theta \longrightarrow 0$, $T_{4a}/T_{44} \longrightarrow 0$, and $\beta_{ab} \longrightarrow$ constant). Collins and Hawking [52] have proved that this is only possible if ${}^3R_{ab}$ approaches isotropy. The suggestion for this arose from the fact that it is clearly true in the cases described by the Lagrangians. The proof requires two steps a) to show that $T^{44}\, e^{(2+\varepsilon)\lambda} \to 0$ for any $\varepsilon < 1$, and b) to use this. The first step follows from the Bianchi identities, and the second from the R_{ab} equations. Both require the dominant-energy and positive-pressure assumptions, which in terms of an orthonormal tetrad, may be expressed as, respectively, $|T^{44}| \geqslant |T^{\alpha}{}_{\beta}|$ and $|T^{a}{}_{a}| \geqslant 0$.

Of the remaining possibilities (namely the group types which include exactly FRW models), type VII_h turns out also to be unstable [52]. The only "general" Bianchi type (in the sense of Section 4.2) remaining is type IX, which does not expand indefinitely but collapses to a second singularity. Thus one can say that in general the long-term effect of the curvature terms of the hypersurface geometry is to

force the ever-expanding models to be anisotropic, in the far future. However [51] they can still be compatible with observation at the present epoch.

5.3. Concluding remarks

To wind up, I would like to go beyond my brief and bring in some more physical considerations.

There are three or four sources of evidence concerning the actual anisotropy of the universe. Direct measurements of the present-day shear give very weak limits [53]. The microwave background isotropy, which is very precise, is determined by the integrated effect over the time since the radiation was last scattered. If the shearing is monotone throughout this period, very strong limits on anisotropy can be derived, of the order of 10^{-3} downwards [54]. I think these could be evaded (if implausibly) by assuming an exact number of cycles of oscillation have occured instead. The third source of evidence comes from primordial element formation. This occurs in a very short time and is highly sensitive to the rate of expansion during that period. This rate itself depends on the amount of shear, and Barrow [55] has estimated, again, assuming monotone behaviour since the element formation epoch limits of the order 10^{-7} or better, cf. [56]. Finally, one may consider the entropy [57]. The idea here is that any dissipative process which removes the anisotropy must simultaneously increase entropy, and from this point of view the observed entropy is remarkably low, implying very low anisotropy at all epochs. There are various physical assumptions in this argument, e.g. that the expansion really is associated with energy in the usual sense and that this energy ends up in the thermal equilibrium represented by the microwave background.

All these considerations suggest that the hypothesis of chaotic cosmology can only be fulfilled if dissipation occurs very early indeed. A theoretical argument shows that the process must be an odd one, if arbitrary anisotropy is to be removed. This is that any system of regular differential equations has a unique solution. So by setting initial data with a large present-day shear, and integrating backwards in time, we could find conditions, at any finite t, giving rise to a model universe incompatible with observation. Thus the processes required should not obey regular differential equations. Of course, the initial big-bang is a point where the governing equations of our models becomes singular, so the most likely possibility is a

process that continues from the initial singularity. This of course must involve quantum mechanics, and here is another speculative area of physics, but one which is being energetically explored.

My own opinion in this is that the various points remarked in this chapter make anisotropic models, except as perturbations of FRW models, implausible.

References

1. Cohn, P.M. 1957, Lie Groups, (Cambridge: University Press)
2. Bianchi, L. 1898, Mem. Soc. It. 11, 267; reprinted in Opere, vol IX, ed. A. Maxia, (Rome: Editzioni Crenonese) (1952), see also 1918, Lezioni sulla teoria dei gruppi continui fini di transformazioni, (Pisa: Edizioni Spoerri).
3. Estabrook, F.B., Wahlquist, H.D. and Behr, C.G. 1968, J. Math. Phys. 9, 497
4. Ellis, G.F.R., MacCallum, M.A.H. 1969, Comm. Math. Phys. 12, 108
5. MacCallum, M.A.H. 1971, Comm. Math. Phys. 20, 57
6. Collins, C.B., Hawking, S.W. 1973, Astrophys. J. 180, 317
7. Siklos, S.T.C. 1978, Comm. Math. Phys. 58, 255
8. Eisenhart, L.P. 1933, Continuous Groups of Tarnsformations, (Princeton: Princeton University Press), reprinted 1961 by (New York: Dover)
9. Schmidt, B.G. 1971, Gen. Rel. Grav. 2, 105
10. Schmidt, B.G. 1968, Riemannsche Raume mit mehrfach transitiver Isometriegruppe, Ph.D. thesis, Hamburg
11. Ryan, M.P., Shepley, L.C. 1975, Homogeneous Relativistic Cosmologies, (Princeton: Princeton University Press)
12. King, A.R., Ellis, G.F.R. 1973, Comm. Math. Phys. 31, 209
13. Ellis, G.F.R., King, A.R. 1974, Comm. Math. Phys. 38, 119
14. Siklos, S.T.C. 1978, Algebraically Special Homogeneous Spacetimes (Oxford University preprint)
15. Newman, E.T., Penrose, R. 1963, J. Math. Phys. 3, 566
16. Kantowski, R. 1966, Some Relativistic Cosmological Models, Ph.D. thesis, (University of Texas at Austin)
17. Collins, C.B. 1977, J. Math. Phys. 18, 2116
18. Kantowski, R., Sachs, R.K. 1966, J. Math. Phys. 7, 443
19. Kompaneets, A.S., Chernov, A.S. 1964, Z.E.T.F. 47, 1939, translation, 1965, Sov. Phys. J.E.T.P. 20, 1303
20. Petrov, A.Z. 1969, Einstein Spaces, (Oxford: Pergamon); cf. 1964, Einstein-Raume, (Berlin: Akademie-Verlag); 1961, Prostranstva Einsteina, (Moscow: Fizmatizd)
21. MacCallum, M.A.H., Stewart, J.M., Schmidt, B.G. 1971, Comm. Math. Phys. 17, 343
22. MacCallum, M.A.H. 1972, Phys. Lett. A40, 385
23. Ehlers, J. 1961, Akad. Wiss. Lit. Mainz., Abh. 11
24. Ellis, G.F.R. 1971, in General Relativity and Cosmology, ed. Sachs, R.K., Proceedings of the Enrico Fermi Summer School, Varenna, Course 47, (New York: Academic Press)
25. Hughston, L.P., Jacobs, K.C. 1970, Astrophys. J. 160, 147
26. Jacobs, K.C. 1977, Source-free Bianchi Electromagnetic Fields, preprint, Max-Planck Institute, Munich, MPI-PAE-Astro 121
27. Misner, C.W. 1968, Astrophys. J. 151, 431
28. Misner, C.W. 1969, Phys. Rev. Letters 22, 1071
29. Misner, C.W. 1969, Phys. Rev. 186, 1319
30. Ozsvath, I. 1971, J. Math. Phys. 12, 1078
31. Hawking, S.W. 1969, Mon. Not. Roy. Astr. Soc. 142, 129

32. MacCallum, M.A.H., Taub, A.H. 1972, Comm. Math. Phys. 25, 173
33. Ryan, M.P. 1974, J. Math. Phys. 15, 812
34. Ryan, M.P. 1972, Hamiltonian Cosmology, (Berlin: Springer-Verlag Lecture Notes in Physics, vol. 10)
35. Sneddon, G.E. 1975, J. Phys. A9, 229
36. Collins, C.B. 1971, Comm. Math. Phys. 23, 137
37. Kubo, M. 1975, Publ. Astr. Soc. Japan 26, 355; and 27, 111
38. Goethals, M. 1975, Ann. Soc. Sci. Bruxelles 89, 50
39. Collins, C.B. 1974, Comm. Math. Phys. 39, 131
40. Shikin, I.S. 1975, Zh.E.T.F. 68, 1583; translated 1975, Sov. Phys. J.E.T.P. 41, 794
41. Collins, C.B. 1972, Comm. Math. Phys. 27, 37.
42. Shikin, I.S. 1972, Sov. Phys. J.E.T.P. 36, 811
43. Belinskii, V.A., Khalatnikov, I.M. 1975, Zh.E.T.P. Pisma 21, 99; and Zh.E.T.P. 69, 401; translated 1975, Sov. Phys. J.E.T.P. 42, 205
44. Bogoyavlenskii, O.I., Novikov, S.P. 1975, in Trudy Seminara I.G. Petrovskovo, Moscow
45. Hawking, S.W., Ellis, G.F.R. 1973, The Large Scale Structure of Space-time (Cambridge: Cambridge University Press)
46. Shepley, L.C. 1969, Phys. Lett. A28, 695
47. Ellis, G.F.R., Schmidt, B.G. 1977, Gen. Rel. Grav. 8, 915
48. Belinskii, V.A., Khalatnikov, I.M., Lifschitz, E.M. 1972, Zh.E.T.P. 62, 1606
49. Moser, A.R., Matzner, R.A., Ryan, M.P. 1973, Ann. Phys. 79, 558
50. Belinskii, V.A., Khalatnikov, I.M., Ryan, M.P. 1972, Ann. Phys. 70, 301
51. Doroshkevich, A.G., Lukash, V.N., Novikov, I.D. 1973, Zh.E.T.P. 64, 1457; translated 1973, Sov. Phys. J.E.T.P. 37, 739
52. Collins, C.B., Hawking, S.W. 1973, Astrophys. J. 180, 317
53. Kristian, J., Sachs, R.K. 1966, Astrophys. J. 147, 379
54. Collins, C.B., Hawking, S.W. 1972, Mon. Not. Roy. Astr. Soc. 162, 307
55. Barrow, J. 1976, Mon. Not. Roy. Astr. Soc. 175, 359
56. Olson, D.W. 1978, Astrophys. J. 219, 77
57. Barrow, J.D., Matzner, R.A. 1978, Mon. Not. Roy. Astr. Soc. 181, 719
58. Kasner, E. 1921, Amer. J. Math. 43, 217
59. Taub, A.H. 1951, Ann. Math. 53, 472
60. Ehlers, J., Kundt, W. 1962, in Gravitation; an introduction to current research, ed. Witten, L., (New York: Wiley)
61. Sanders, P.T. 1967, Ph.D. thesis, King's College, London
62. Heckmann, O., Schucking, E. 1962, in Gravitation; an introduction to current research, ed. Witten, L. (New York: Wiley)
63. Robinson, B.B. 1961, Proc. Nat. Acad. Sci. 47, 1852
64. Doroshkevich, A.G. 1965, Astrophysics 1, 138
65. Shikin, I.S. 1966, Doklady Akad. Nauk, USSR, 171, 73
66. Thorne, K.S. 1967, Astrophys. J. 148, 51
67. Stewart, J.M., Ellis, G.F.R. 1968, J. Math. Phys. 9, 1072
68. Rosen, G. 1962, J. Math. Phys. 3, 313
69. Jacobs, K.C. 1969, Astrophys. J. 155, 379
70. Newman, E.T., Tamburino, L.A., Unti, T. 1963, J. Math. Phys. 4, 915
71. Carter, B. 1968, Phys. Lett. A26, 399
72. Carter, B. 1968, Comm. Math. Phys. 10, 268
73. Cahen, M., Defrise, L. 1968, Comm. Math. Phys. 11, 56
74. Maartens, R., Nel, S.D. 1977, Decomposable Differential Operators in a Cosmological Context, preprint, Cape Town
75. Farnsworth, D. 1967, J. Math. Phys. 8, 2315
76. Wainwright, J., Ince, W.C.W., Marshnan, B.J. 1978, Spatially Homogeneous Cosmologies with Equation of State p , Univ. of Waterloo, preprint

77. Melvin, M.A. 1975, Ann. N.Y. Acad. Sci. 262, 253
78. Brill, D. 1965, Phys. Rev. 133, B845
79. Ozsvath, I. 1966, in Essays in Honour of V. Hlavaty (Bloomington: Indiana University Press)
80. Raychaudhuri, A.K. 1958, Proc. Phys. Soc. 73, 263
81. Jacobs, K.C. 1968, Astrophys. J. 153, 66
82. Rosen, G. 1964, Phys. Rev. 136, B297
83. Ozsvath, I. 1977, Gen. Rel. Grav. 8, 737
84. Barnes, A. 1978, J. Phys. A11, 1303
85. Collins, C.B. 1977, J. Math. Phys. 18, 2116
86. Harvey, A., Tsoubelis, D. 1977, Phys. Rev. D15, 2734
87. Hughston, L.P., Shepley, L.C. 1970, Astrophys. J. 160, 333
88. Dunn, K.A., Tupper, B.O.J. 1976, Astrophys. J. 204, 322
89. Collinson, C.D., French, D.C. 1967, J. Math. Phys. 8, 701
90. Demianski, M., Grishchuk, L.P. 1972, Comm. Math. Phys. 25, 233
91. Lukash, V.N. 1974, Zh.E.T.P. 67, 1594; translated 1975, Sov. Phys. J.E.T.P. 40, 792
92. Barrow, J.D. 1976, Mon. Not. Roy. Astr. Soc. 175, 359
93. Bondi, H. 1960, Cosmology, (Cambridge: Cambridge University Press)
94. Will, C.M. 1974, in Proceedings of the Enrico Fermi Summer School, Varenna, Course 56, New York: Academic Press)
95. Will, C.M. 1979, in Gravitational Theory Since Einstein ed. Hawking, S.W., Israel, W., (Cambridge: Cambridge University Press)
96. MacCallum, M.A.H. 1973, in Cargese Lectures, vol. 6, ed. Schatzmann, E., (New York: Gordon and Breach)
97. MacCallum, M.A.H. 1979, in Gravitation Theory Since Einstein ed. Hawking, S.W., Israel, W., (Cambridge: Cambridge University Press)
98. Ruffini, R., (ed.), 1979, Bianchi Cosmologies, to appear
99. Misner, C.W. 1969, in Colloques de C.N.R.S. 170, 155
100. Belinskii, V.A., Khalatnikov, I.M., Lifschitz, E.M. 1970, Adv. Phys. 19, 525
1969, Uspekhi Fiz. Nauk 102, 463; translated 1970, Sov. Phys. Uspekhi 13, 745
101. de Vaucouleurs, G. 1970, Science 167, 1203
102. Ellis, R.S., Fong, R., Phillipps, S. 1976, Mon. Not. Roy. Astr. Soc. 176, 391
103. Soneira, R.M., Peebles, P.J.E. 1976, Astrophys. J. 211, 1
104. Turner, E.L., Gott III, J.R. 1974, Astrophys. J. 197, L89
105. de Vaucouleurs, G. 1976, Astrophys. J. 205, 13
106. Rubin, V.C., Thonnard, N., Ford, W.K., Roberts, M.S. 1976, Astron. J. 81, 719
107. Schechter, P.L. 1977, Astron. J. 82, 569
108. Jaakkola, T., Karoji, H., Le Dermat, G., Moles, M., Nottale, L., Vigier, J.P., Pecker, J.C. 1976, Mon. Not. Roy. Astr. Soc. 177, 191
109. Webster, A. 1976, Mon. Not. Roy. Astron. Soc. 175, 61 and 175,71
1977, Mon. Not. Roy. Astron. Soc. 179, 511 and 179, 517 (with Pearson, T.J.)
110. Silk, J. 1971, Space Science Rev. 11, 671
111. Peebles, P.J.E. 1971, Physical Cosmology, (Princeton: Princeton University Press)
112. Smooth, G.F., Gorenstain, M.V., Muller, R.A. 1977, Phys. Rev. Lett. 39, 898
113. Kobolov, V.M., Reinhardt, M., Sazonov, V.N. 1976, Astrophys. Lett. 17, 185
114. Rozmaikin, A.A., Sokoloff, D.D. 1977, Astron. and Astrophys. 58, 247
115. Reinhadrt, M. 1971, Astrophys. and Sp. Sci. 10, 363
116. Brown, F.G. 1968, Mon. Not. Roy. Astron. Soc. 138, 527
117. Reinhardt, M. 1971, Mon. Not. Roy. Astron. Soc. 156, 151
118. Hauley, D.L., Peebles, P.J.E. 1975, Astron. J. 80, 477

119. Fesenko, B.I. 1976, Astron. Zh. 53, 1153; translated 1977, Sov. Astr. A.J. 20, 550
120. Thompson, L.A. 1973, Publ. Astr. Soc. Pacific 85, 528
121. Thompson, L.A. 1976, Astrophys. J. 209, 22
122. Borchkhadze, T.M., Kogoshvili, N.G. 1976, Astron. and Astrophys. 53, 431
123. Willson, M.A.G. 1972, Mon. Not. Roy. Astr. Soc. 155, 275
124. Joseph, K. 1966, Proc. Camb. Phil. Soc. 62, 87.

CREATION OF PARTICLES BY GRAVITATIONAL FIELD

Ya.B. Zel'dovich
Institute of Applied Mathematics, Academy of Sciences USSR, Moscow

I would like to review recent work on the problem of creation of particles by gravitational field but without going into all the technical details. Everybody who wants to work on this problem should read original papers (see reviews Parker (1977) and (1979), Isham, Penrose, Sciama (1975), Starobinsky, Zel'dovich (1979)). In my lecture I will concentrate on general principles leaving out many formulas and derivations.

We shall use quantum theory to describe particles (electrons, neutrinos etc.) and fields, for example scalar and electromagnetic field, in a classical space-time. This means that we assume that the gravitational field or in other words that the metric of space-time is well defined. It does not imply however that we know the metric exactly. The back reaction of created particles can deform space-time in a complicated way (Wald (1977)) but let us assume that we know the metric as precisely as is necessary and the metric is not restricted by any uncertainty principle. Uncertainty principles apply only to particles and fields. In other words the gravitational field is treated classically, it is not quantized. One can go a step further and decompose the metric into a smooth part connected with matter and an oscillating part describing gravitational waves. Gravitational waves can be quantized and in natural way gravitons will emerge. In this way we quantize small perturbations of the background metric treated as a classical field. The new and very interesting possibility is connected with excitations leading to changes in topology, they cannot be treated by perturbation theory. They appear only in nonlinear theories and general relativity is indeed a nonlinear theory. In the classical case these excitations are represented by particle like solutions (solitons), in the quantum field theory they are represented by solitons in a non-physical Euclidean space-time (with imaginary time) and they describe tunneling between degenerate vacua. In that case, they are called instantons. Unfortunately here I will not be able to discuss these new developments. We will consider only linear field theory on a curved space-time.

Theory of quantum creation of particles by a classical gravitational field can be formulated in three steps. In the first step one

studies motion of point-like test particles in a given metric. Test particles move along geodesics so one has to study geodesics in space-time. In the second step one formulates a theory of classical fields in a given space-time, for example one considers a classical electromagnetic field in a curved space-time. Finally in the third step one considers the quantum field theory on a given curved background, for example one can quantize an electromagnetic field in a given metric (Sexl and Urbantke (1967)). It turns out that the mathematical basis needed for the quantum description of fields is almost the same as that needed for classical field theory. The final step is easy and formulating quantum field theory one is guided by fundamental principles but they should be always clearly stated and well understood.

Let us begin with the second step, and consider classical electromagnetic field in a given space-time. We will take a metric of the following type: for $t < t_1$ and $t > t_2$ (with $t_2 > t_1$) the space-time is flat and is described by the Minkowski metric, and for $t_1 < t < t_2$ the metric is perturbed and non flat. We assume that we know the classical solution for $t < t_1$ and we denote it by $\vec{E}_{in}$ and $\vec{H}_{in}$. In the region $t_1 < t < t_2$ we have some fields $\vec{E}$ and $\vec{H}$, and for $t > t_2$ we have $\vec{E}_{out}$ and $\vec{H}_{out}$. The "in" fields can be decomposed into simple harmonic waves. We cannot do that with the fields $\vec{E}$ and $\vec{H}$. If we know the value of fields at $t = t_2$ then we also can decompose the "out" fields into harmonic waves. The classical theory of electromagnetic field is linear, therefore we are sure that if we multiply the "in" fields by a constant a then the "out" fields should be also multiplied by a. So if initially we had vacuum (a = 0) then in the classical theory we get the vacuum in the out state. In the classical theory particles are not created from the vacuum. On the other hand we have the correspondence principle, which states that quantum theory for large values of occupation numbers (a strong electromagnetic field) goes over into the classical theory. Let us assume that the curved metric is such, that the outgoing field is stronger than the incoming field. If initially the number of particles or photons was n then finally we can observe another number of particles, so the classical theory tells us that creation of particles is possible but not from vacuum, one needs for that non-zero initial fields. As long as we are in the framework of classical theory we know that, normally, in flat space, creation of photons is possible when some charges oscillate but in gravitational field new photons could be created even if charges do not oscillate. In this case the rate of creation of pairs of photons is proportional to the number density of already existing photons.

This is well known and it reminds the basic principle of lasers.

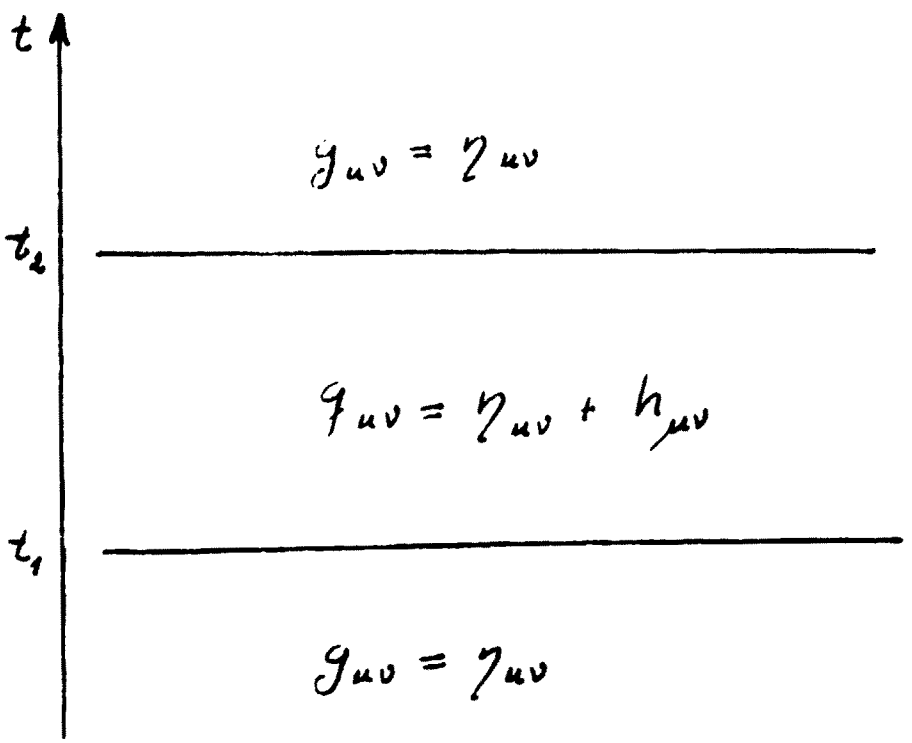

Fig. 1. Sandwich type space-time with Minkowski space-time at $t < t_1$ and $t > t_2$ ($t_2 > t_1$) and perturbed (curved) space time for $t_1 < t < t_2$

Now let us go to the quantum theory. In quantum theory $\vec{E}$ and $\vec{H}$ are represented by operators, and if you consider the scalar field $\varphi(t,\vec{x})$ then φ and $\dot{\varphi}$ are operators just as in quantum mechanics position x of a particle and its momentum $p = m\dot{x}$ are operators. Quantum field theory is similar to the theory of many particles with coordinates of a particle corresponding to a value of the field at a given point. I am not going to discuss the formal structure of quantum field theory (see Schweber (1961), Bjorken, Drell (1965), Bogo-

liubov, Shirkov (1959)). Let me only recall that the overall energy of a scalar field or electromagnetic field is given by

$$\int T_{oo}\, dV = \sum_k (n_k + \tfrac{1}{2})\, \hbar\, \omega_k \qquad (1)$$

and different occupation numbers n_k correspond to different states of the field. Now one can write

$$T_{oo} = \sum_k (n_k + \tfrac{1}{2})\, \hbar\omega_k, \quad T_{ij} = \sum_k (n_k + \tfrac{1}{2})\, \frac{\hbar k_i k_j}{\omega_k} \qquad (2)$$

where k_i is a wave vector of the k-th mode. Therefore as a result of introduction of operators and application of rules of quantum field theory we obtain the particle picture of the field. Energy of the field is a sum of energies of particles and the energy momentum tensor is a sum of contributions of particles. Existence of photons and particles is a natural result of the quantization procedure - of introduction of operators and commutation relations of fields in the Minkowski space.

We obtained the notion of a "photon as a particle" and a particle picture of the electromagnetic field. Different occupation numbers (numbers of photon) correspond to different states of the field. The vacuum is by definition the lowest energy state, therefore the vacuum corresponds to all $n_k = 0$. We see that quantum field theory leads to a strange result: the energy density of vacuum is not zero. It is lowest, it is minimum in comparison with other states but it is not zero. Energy of the vacuum is given by $\sum_k \frac{1}{2}\hbar\omega_k$ and since in general number of states is infinite it is a divergent quantity. This fact is harmless in laboratory physics, where one can neglect effects of gravitational field. In laboratory we always have to deal with differences of energies of two states. For example, initially we have an excited atom with some excitation energy and infinite energy of the vacuum, then the atom jumps to its ground state emitting a photon so the final energy is an energy of the atom plus the energy of the photon plus infinite energy of the vacuum and practically we can forget about the infinite energy of the vacuum. The situation is different when gravitation is taken into account. The energy momentum tensor appears on the right hand side of the Einstein equations and we cannot put infinity there (Adler, Liebermann and Ng (1977), Bernard and Duncan (1977), Davies et al. (1977)).

There are at least two possible ways of approaching this problem. One can modify the expression for the energy momentum tensor

and hence the rules of working with operators so as to set the energy of the vacuum ϵ_v to be equal to zero. If you adopt this convention then you will find that $\langle T_{oo} \rangle \neq \langle E^2 + H^2 \rangle$ and moreover in the vacuum it is not possible to have simultaneously $\langle E \rangle = 0$ and $\langle H \rangle = 0$ in all states, because this will contradict the Heisenberg uncertainty relation. So when $\epsilon_v = 0$, $\langle E^2 \rangle$ and $\langle H^2 \rangle$ can not vanish. This is due to ordering convention of operators.

There is another approach which is recently very popular and which is related to the most important problem in physics. The total vacuum energy density is equal to sum of vacuum energy densities of different fields. Moreover the vacuum energy momentum tensor has to be proportional to the metric tensor in a Lorentz invariant theory, and therefore it could be identified with the cosmological constant term. The most important restriction imposed on the quantum theory is the requirement that the cosmological constant has to be equal to zero or to be very small, according to the present observational data. I do not know of any theory which satisfies this requirement. One possible approach is to reduce to zero separately the energy of the vacuum of electromagnetic field and all other fields present. There is still another possibility namely to assume that contributions of different fields to the energy of the vacuum are of different signs and that the sum of all contributions is equal to zero, but the vacuum energy of every field is equal to plus or minus infinity. Supersymmetry theory is an example of a theory realizing this program (Golfand and Likhtman (1971), Akulov and Volkov (1973), Wess and Zumino (1974)). In supersymmetric theory you have two fields a spin two boson field representing gravitons and a spin 3/2 fermion field representing new not yet observed particles (Deser and Zumino (1976), Freedman, van Nieuwenhuizen and Ferrara (1976)). These fields exist and interact in such a way that the total energy of the vacuum is equal zero and also other normally diverging quantities are finite. Therefore though the energy of the vacuum for a single field is infinite the total zero point (vacuum) energy is equal to zero. Particles of spin 3/2 appearing in the supersymmetric theory have not yet been found in nature, therefore it is premature to say that it is the right theory. To really realize the whole program it would be necessary to include all fields found in nature which are responsible for strong, weak, electromagnetic and gravitational interactions. This was never done and it is only a dream of theoreticians. Existence of instantons makes this problem even more difficult.

Situation is even more complicated in curved space since in gen-

eral the curvature will also give some contribution to the zero point energy and it at least should be renormalizable to a finite contribution (DeWitt (1975), Ford (1975), (1976)). We shall come back to this point later. At the moment let us only remark that renormalization procedures could at most reduce the zero point energy to zero, but none of them could reduce simultaneously $\vec{E}$ and $\vec{H}$ of the vacuum to zero. $\vec{E}$ and $\vec{H}$ of the vacuum are observed in Lamb shift measurements.

It is quite easy to describe creation of particles. If for $t < t_1$ we have vacuum, then at $t > t_2 > t_1$ we can have some particles. Let us concentrate on a definite mode of the field $(E_k, H_k)_{in}$ with a wave vector $\vec{k}$ and vacuum energy $\frac{1}{2}\hbar\omega_k$ and let the outgoing field be such that $(T_{\mu\nu})_{out} = 1{,}2\ (T_{\mu\nu})_{in}$ (classical) so that the final energy is $0{,}6\,\hbar\omega_k$. The net gain in energy is therefore $0{,}1\hbar\omega_k$, this energy is created. One can think that something is wrong since the number of created photons is not integer. Our result is correct; it means that the probability of creation of a photon is 0,1. Therefore once you have a classical description of the field you can go over into the quantum formulation and the zero point oscillations correspond to the initial field, which could be amplified. So we obtain not only creation of photons from already existing photons but also creation of photons from vacuum. This is similar to the situation in quantum optics where if the system emits induced radiation, then you know that it can also emit spontaneous radiation (Einstein (1917)). Creation of particles from vacuum is in the some relation to the classical situation as the spontaneous radiation of atoms is to the induced radiation.

In this simple description we are facing a certain problem. The classical electromagnetic theory is invariant with respect to time reversal and you can change the arrow of time. If from $(T_{\mu\nu})_{in}$ at $t < t_1$ we obtain $1{,}2\ (T_{\mu\nu})_{in}$ at $t > t_2$ then taking the reverse direction of time ($t \longrightarrow -t$) we would get that from $1{,}2\ (T_{\mu\nu})_{in}$ we obtain finally only $(T_{\mu\nu})_{in}$ or from $(T_{\mu\nu})_{in} \longrightarrow 0.8\ (T_{\mu\nu})_{in}$, which is less than in the initial state, but this is of course impossible. The point is that the quantities which we are calculating are already averaged over phases. A classical field with a given phase could be amplified or damped when it passes through curved space-time ($t_1 < t < t_2$). To specify the classical electromagnetic field one has to specify also its phase but then the field has an undetermined number of photons. A field with exactly determined number of photons has undetermined phase. There is uncertainty relation between phase and occupation number just as between coordinate and momentum. Therefore the vacuum cor-

responds to a field with undetermined phase (phase is defined for standing waves) but for $t > t_2$ we have a field with given energy and given momentum and with partly determined phase. We average over phases of the initial situation and this introduces asymmetry. After averaging over phases of the initial situation we always obtain a gain in energy in the final state. Approximately we have the following situation: if initially the amplitude of the field is A_{in} then the final amplitude A_{out} is

$$A_{out} = A_{in} (1 + h \cos \Theta), \tag{3}$$

where h describes small departures from the flat metric, so we have

$$g_{\mu\nu} = \eta_{\mu\nu} + h q_{\mu\nu}(t), \tag{4}$$

and $q_{\mu\nu} = 0$ for $t \longrightarrow \pm\infty$. If we take square of the final amplitude then we get

$$A^2_{out} = A^2_{in} (1 + 2h \cos \Theta + h^2 \cos^2 \Theta), \tag{5}$$

and averaging over Θ we obtain

$$\langle A^2_{out} \rangle = \langle A^2_{in} \rangle (1 + \tfrac{1}{2} h^2), \tag{6}$$

so we get a positive contribution. In the reverse situation it is necessary to average first over phases at the other end of the process. Asymmetry is connected with the fact that we have to take an average over phases of the initial situation. As a partial result we notice that the rate of creation of particles is proportional to h^2 and not to h (Starobinsky and Zel'dovich (1972), Woodhouse (1976)).

Let us now return to the classical calculations and instead of working with electromagnetic field, which has many components, we will restrict ourselves to a scalar field which satisfies the wave equation

$$\Box\varphi = \frac{\partial^2 \varphi}{\partial t^2} - \Delta\varphi = 0 \tag{7}$$

where units are such that c = 1. The metric of the background space-time we take in the form

$$ds^2 = dt^2 - a^2(t)dx^2 - b^2(t)dy^2 - c^2(t)dz^2 \tag{8}$$

i.e. with flat three-dimensional spatial sections t = const. Assuming that

$$\varphi = f(t)\ e^{i\vec{k}\vec{x}}, \tag{9}$$

the wave equation reduces to

$$\frac{d^2 f}{dt^2} = - \omega^2 f, \tag{10}$$

where

$$k_x^2 a^{-2} + k_y^2 b^{-2} + k_z^2 c^{-2} = \omega^2(t). \tag{11}$$

The general solution of (10) when ω = const is

$$f = A e^{-i\omega t} + B e^{i\omega t} \tag{12}$$

(For one running wave $A = A_o$, $B = 0$ one cannot define the phase by taking $e^{-i\omega t + i\theta}$ instead of $e^{-i\omega t}$ since it is equivalent to spatial displacement. The phase is important as a relation between A_o and B_o or between $A_o(k)$ and $A_o(-k)$ which brings us back to standing waves. The calculation for one running wave $A = A_o$, $B = 0$ is already equivalent to the phase-averaged calculation for standing waves.) In this simple case of homogeneous metric no mixing of modes occurs. One can now work with a single mode and we have a situation similar to mechanical oscillations with time dependent period (pendelum with slowly varying length).

As is well known if ω is slowly varying in time we can use the adiabatic approximation and replace $f = A e^{-i\omega t}$ by

$$f = A e^{-i\int \omega \, dt}. \tag{13}$$

Allowing A to be time dependent (10) leads to

$$0 = \ddot{f} + \omega^2 f = -2i\omega \frac{\dot{A}}{A} f - i\dot{\omega} f + \left(\frac{\dot{A}}{A}\right)\dot{} f + \left(\frac{\dot{A}}{A}\right)^2 f, \tag{14}$$

and neglecting quadratic terms and higher derivatives we get

$$\frac{\dot{A}}{A} = - \frac{1}{2} \frac{\dot{\omega}}{\omega} \tag{15}$$

This equation can be easily solved and finally we obtain

$$A = \frac{\alpha}{\sqrt{\omega}}, \tag{16}$$

Therefore the general solution of equation (10) in the adiabatic approximation is

$$f = \frac{\alpha}{\sqrt{\omega}} e^{-i\int \omega \, dt} + \frac{\beta}{\sqrt{\omega}} e^{i\int \omega dt} \tag{17}$$

where α and β are constants but $\omega = \omega(t)$. Now we can go further to the post adiabatic approximation and assume that α and β

depend on time. In this case the decomposition of the field into positive and negative frequencies is not unique. We can use this freedom to impose additional conditions, which are necessary to get rid of some pathological situations. This method of solving equation (10) was invented by Lagrange.

Assuming that the first time derivatives of α, β and ω are small we can compute the first time derivative of f and get

$$\dot{f} = -i\alpha\sqrt{\omega}\,e^{-i\int\omega dt} + i\beta\sqrt{\omega}\,e^{i\int\omega dt} \tag{18}$$

This relation is true in the general case if

$$(\dot{\alpha} - \frac{1}{2}\frac{\dot{\omega}}{\omega}\alpha)e^{-i\int\omega dt} + (\dot{\beta} - \frac{1}{2}\frac{\dot{\omega}}{\omega}\beta)e^{i\int\omega dt} = 0 \tag{19}$$

Equation (10) will be satisfied if in addition to (19) we have

$$-i\alpha\sqrt{\omega}(\frac{\dot{\alpha}}{\alpha} + \frac{1}{2}\frac{\dot{\omega}}{\omega})e^{-i\int\omega dt} + i\beta\sqrt{\omega}(\frac{\dot{\beta}}{\beta} + \frac{1}{2}\frac{\dot{\omega}}{\omega})e^{i\int\omega dt} = 0 \tag{20}$$

From (19) and (20) we can compute $\dot{\alpha}$ and $\dot{\beta}$

$$\begin{aligned}\dot{\alpha} &= \frac{1}{2}\frac{\dot{\omega}}{\omega}\beta\,e^{2i\int\omega dt}\\ \dot{\beta} &= \frac{1}{2}\frac{\dot{\omega}}{\omega}\alpha\,e^{-2i\int\omega dt}\end{aligned} \tag{21}$$

This system of coupled first order differential equations has a first integral

$$|\alpha|^2 - |\beta|^2 = \text{const} \tag{22}$$

To solve equations (21) we take as initial conditions $\alpha_{in} = 1$, $\beta_{in} = 0$, then we get in the first approximation

$$\beta_{out} = \frac{1}{2}\int \frac{\dot{\omega}}{\omega}\,e^{-2i\int\omega dt}\,dt \tag{23}$$

Let the metric coefficients be such that

$$a,b,c \sim (1 + hq(t)) \tag{24}$$

then using (11) we get

$$\omega = \omega_0(1 + 2hq(t)) \tag{25}$$

where q(t) is such that $q(t) \xrightarrow[t\to\pm\infty]{} 0$. The function q(t) is normalized for example, $q_{max} = 1$, h denotes a small parameter and higher powers of h are neglected. With this additional information we can insert ω into (23) and obtain in the first approximation

$$\beta_{out} = \frac{1}{2}\int \frac{\dot{\omega}}{\omega} e^{-2i\int\omega dt}\, dt = h\int \dot{q} e^{-2i\omega_o t}\, dt = h\dot{q}_{2\omega_o} \qquad (26)$$

so $|\beta_{out}|^2 = h^2 |\, q_{2\omega_o}|^2$ and therefore $|\alpha_{out}|^2 = 1 + h^2 |\, \dot{q}_{2\omega_o}|^2$.

What we can learn from this result? In the classical case the momentum is fixed and we have two frequencies positive and negative corresponding to waves traveling in opposite directions. When initially we had only one wave propagating in a given fixed direction, then a wave propagating in opposite direction is created and the initial wave is amplified, which means that a pair of particles with opposite momenta is created. Therefore particles are created in pairs and they are moving in opposite directions. Momentum is conserved and this is natural because, due to symmetries of the background space-time, momentum has to be conserved. Since particles are created in pairs, in the case of electromagnetic field we have to create simultaneously two photons and for this we need a portion of energy equal to $2\cdot\hbar\omega_o$. This energy is taken from the gravitational field and that is why in (26) we have $\dot{q}_{2\omega_o}$. Density of created particles is given by

$$N\,[cm^{-3}] = h^2 \int |\dot{q}_{2\omega}|^2 k^2\, dk, \qquad (27)$$

which for massless particles reduces to

$$N = h^2 \int |\dot{q}_{2\omega}|^2 \omega^2\, d\omega = h^2 \int |\ddot{q}_{2\omega}|^2\, d\omega = h^2 \int |\ddot{q}|^2\, dt, \qquad (28)$$

so we have

$$\frac{dN}{dt} = \frac{1}{c^3} h^2 |\ddot{q}|^2 \qquad (29)$$

where we have introduced the velocity of light c and now it is apparent that our result is correct from the point of view of dimensions (h and q are dimensionless, the dimension of $\ddot{q}$ is t^{-2}). Energy density of created particles is given by

$$\epsilon_{out} = h^2 \int |\dot{q}_{2\omega}|^2 \hbar\omega\omega^2\, d\omega \qquad (30)$$

If the perturbation of metric is described by a smooth function q(t) then all quantities connected with the outgoing field are regular. We exclude therefore δ functions, θ functions and functions with non continuous first derivatives. When q is smooth and has no singularities in the complex plain, then for large ω, $\dot{q}_{2\omega}$ is exponentially decreasing and we get finite result at $t \to \infty$.

Up to now we discussed boson fields. Particles with integer spins-bosons, result directly from the application of quantum theory

to the classical field theory. When fields are c numbers one can describe them by the classical Lagrangian but if fields are represented by operators then they in a natural way lead to particles, for example photons. In the theory of fermion fields one begins with particles. To describe particles we can use Schrödinger equation or, if particles are relativistic, Dirac equation. Dirac equation predicts that there should be positive and negative energy states. From experimental data we know that the wave function for a system of many fermions is antisymmetric. Dirac equation forces us to accept the fact that there is a sea of particles with negative energies and the Pauli exclusion principle tells us that in vacuum the sea is always filled. A hole left by a particle in a ses is observed as an antiparticle. I would like to point out that the believe that the Pauli exclusion principle can be deduced from the relativistic quantum field theory is not quite correct. Without the Pauli principle the theory will not be inconsistent but it will be in contradiction with simple observations. For example, in that case atoms could not exist. Particles would then jump freely from positive to negative energy states (without creation of antiparticles) and this is not observed. But with the Pauli principle the Dirac theory describes processes which really exist and are observed. In the Dirac theory an electron can jump from one positive energy state to another positive energy state. This takes place in atoms and it is a description of motion of an electron. There are also jumps from negative energy states to positive energy states and these describe creation of a pair of particle and antiparticle. Following the rules of the Dirac theory one can make many calculations, also in curved space-time, and in particular one can treat creation of particles as a kind of motion. In curved space-time a quantum particle can acquire energy, can change momentum and in particular can change energy state and jump from negative energy state to a positive energy state. At first glance, creation of particles in the Dirac theory has nothing to do with the picture of creation of particles in electromagnetic field but in fact it is very similar. In the electromagnetic case particles appear in pairs and in the Dirac theory they also are created in pairs since an electron jumping from a negative state to a positive energy state leaves a hole in the sea of negative states which is interpreted as antiparticle. In the Dirac theory it is obvious that energy needed to creat a pair of particles is at least equal to $2mc^2$. In electromagnetic theory we have the same situation, we need $2\hbar\omega$ for two photons with frequency ω to be created. So though the initial principles are very differ-

ent it is well known that the theory of fermions and bosons is quite similar.

Let us repeat the general properties of the quantum creation of particles in gravitational field. Particles are created in pairs. For charged particles it is so because charge has to be conserved, neutral fermions are created in pairs because a particle jumping from a negative energy state to a positive energy state leaves a hole and neutral bosons are created in pairs because momentum has to be conserved in spatially uniform space-time. But in fact in every case particles are created in pairs because the gravitational field is coupled with other fields through energy momentum tensor, which is constructed from quadratic expressions in these fields. In spatially nonhomogeneous case particles created in a pair not necessarily have to go in opposite directions.

With a special attention one should treat the case of massless particles. For massless particles the theory can be conformally invariant, it means that if we have a metric

$$\tilde{ds}^2 = A^2(t,\vec{x}) \left[dt^2 - dx^2 - dy^2 - dz^2 \right] \tag{31}$$

then solution of Maxwell equations and other wave equations in this space-time can be obtained from solutions in Minkowski space. For example, for the electromagnetic field we have

$$\tilde{F}_{\mu\nu} = F_{\mu\nu}(\text{Minkowski}) \tag{32}$$

and for spin 1/2 massless field (neutrinos)

$$\tilde{\psi} = A^{-3/2} \psi \quad (\text{Minkowski}) \tag{33}$$

Obviously in Minkowski space it is not possible to creat particles from vacuum. Therefore massless particles are not created in a conformally flat space-time. In geometrical optics approximation massless particles in Minkowski space move along null lines so $ds = 0$ and therefore also $\tilde{ds} = 0$. Particles, which move in Minkowski space along null lines, in the conformally related space also move along null lines. Therefore neutrinos and photons are not created in a conformally flat space-time. Rate of creation of electrons and other massive particles depends on their masses. When electrons are ultrarelativistic then the rest mass contribution to total energy could be considered as a small perturbation and it is easy to show that the rate of creation of ultrarelativistic particles in conformally flat space-time is proportional to m^2 and therefore it is very small.

Let us now give some numerical estimates. If metric perturbations are given by $h_o e^{-t^2/\tau^2}$ or $h_o \tau^2/t^2 + \tau^2$ so that characteristic time scale of perturbation is τ and characteristic amplitude is h_o then the frequency excited is of the order of $\omega \sim \tau^{-1}$ and the volume occupied in phase space is $\omega^3 \sim \tau^{-3}$. Therefore energy density of created particles is

$$h_o^2 \hbar \omega \omega^3 \sim h_o^2 \frac{\hbar}{\tau} \frac{1}{\tau^3} \sim \frac{1}{\tau^4} \tag{34}$$

so it strongly depends on characteristic time scale of perturbation. We can also estimate the energy density needed to influence the background space-time. Curvature is roughly given by $G\ddot{h} \sim \frac{Gh_o}{\tau^2}$ so

$$\frac{\text{energy density created}}{\text{energy density needed}} \sim h_o \frac{t_{Pl}^2}{\tau^2} \tag{35}$$

where $t_{Pl} = (G\hbar/c^5)^{1/2} = 5.4 \ \ 10^{-44}$ s. When $h_o \ t_{Pl}^2/\tau^2$ is of the order of unity then the influence of created particles on the background space-time has to be taken into account.

Up to now we have considered only the following situation: we took the vacuum to be the "in" state. Then we introduced a small perturbation and after some time we switched it off. In the "out" state after subtracting the zero point energy (the energy of vacuum) we noticed that a finite energy density was produced. But if we would be interested in the situation in curved space-time (at a moment when perturbation was not yet switched off) then it turns out that it is not enough to subtract the zero point energy. In Minkowski space the zero point energy ϵ_v diverges as a fourth power

$$\epsilon_v = \int_0^\Lambda \frac{1}{2} \hbar \omega \omega^2 \, d\omega \sim \Lambda^4 \tag{36}$$

but in perturbed space-time additional term proportional to $\ddot{q} \int \omega d\omega$ appears. Renormalization in this case is more then just subtraction of zero point energy, it is also necessary to adjust the gravitational constant. I shall not go into more details here. The important point is that if in the initial state in Minkowski space there are no particles (vacuum), then after perturbation is switched off we would discover that energy density proportional to h^2 was produced. It is quite understandable since interaction is proportional to h and therefore matrix elements giving us transition probabilities are proportional to h^2. In the curved region of space-time we have also vacuum polarization, something unrelated to particles and this effect is proportional to the first power of h, the stresses T_{ij} $(i,j = 1,2,3)$ are proportional to h while energy density $T_{oo} \sim h^2$. It is not surprising

from the point of view of the law of energy conservation that in order to obtain energy which depends quadratically on perturbation one has to have stresses which depend linearly on perturbation. For normal particles pressure p can not be larger than energy density, we always have $T_{oo} \geqslant T_{ij}$ and this relation is called the energy dominance principle. It was shown by Hawking that if the energy dominance principle holds, creation of particles is impossible (Hawking (1970)). In the real process of creation of particles in curved space-time, when metric is changing with time, the energy dominance principle is violated. First of all it means that it is not possible to describe the vacuum polarization using a particle picture. The fact that the energy dominance principle could be violated by quantum processes is very important. There are very general theorems on the existence of singularities which relay on this principle (Hawking and Penrose (1970)). It is quite possible that on the quantum level all or some of the singularities could be avoided. One has to remember however that quantum effects play a dominant role in very special and extreme conditions.

Let me make one more comment. Any conformally invariant theory represents massless particles. The converse statement is not true however. For example, the scalar wave equation can be written in a conformally invariant form but it also could be taken in a non conformally invariant form and actually only experimental tests could distinguish between these two possibilities. As was pointed out by Grishchuk (Grishchuk (1974)) only gravitons are exceptional. It turns out that wave equation for gravitons is uniquely determined by Einstein equations and this equation is not conformally invariant. This is important because the Friedman-Lemaître metric

$$ds^2 = dt^2 - a^2(t)\left[dx^2 + dy^2 + dz^2\right] = a^2(\tau)\left[dt^2/a^2(t) - dx^2 - dy^2 - dz^2\right] \qquad (37)$$

is conformally flat and creation of massless particles except gravitons is not possible in this space-time. It means that overall isotropic expansion does not lead to creation of massless particles except gravitons. In the case of normal matter, using the language of continuous mechanics one would say that second viscosity of the vacuum vanish. Second viscosity is connected with that part of energy momentum tensor which arises from isotropic expansion. So the vacuum of massless conformally invariant particles has no second viscosity. Particles which have mass - electrons, protons etc. - are created in Friedman-Lemaître universe when $t \sim \hbar/mc^2$ so when characteristic time of expansion of

universe is close to their "Compton time". Particles are not created before because then they are ultrarelativistic and effectively conformally invariant and they are not created after because the frequency of gravitational field is then much lower than Compton frequency of particles. So massive particles are created only in a definite period of isotropic expansion. Characteristic parameter which tells when it is necessary to take into account the back reaction of created particles is now $Gm^2/\hbar c$ and for all known particles in nature this parameter is very close to zero. If universe is quasi isotropic then it is not necessarily conformally flat and its curvature is determined by space derivatives of the metric as well as by time derivatives but terms involving spatial derivatives are near the singularity of lower order and therefore the rate of creation of massless particles is again small (Zel'dovich (1973)). Gravitons are not created if the equation of state is $p = \frac{1}{3}\epsilon$. This is the case when scalar curvature R is equal to zero. If $p < \frac{1}{3}\epsilon$ gravitons are created but not very efficiently and since during expansion of universe their energy density decreases faster than energy density of other particles, they are not dynamically important. Only when $p > \frac{1}{3}\epsilon$ creation of gravitons is effective but I do not think that equations of state for which $p > \frac{1}{3}\epsilon$ are realistic. We conclude therefore that in general in the case of quasi isotropic expansion of universe creation of particles does not play an important role.

The most interesting from the cosmological point of view is creation of particles in anisotropic universes. In our group we have investigated creation of particles in Kasner universe and other homogeneous but anisotropic world models(Starobinsky and Zel'dovich (1972), Lukash and Starobinsky (1974)). The Kasner metric is usually written in the form

$$ds^2 = dt^2 - t^{2p_1}dx^2 - t^{2p_2}dy^2 - t^{2p_3}dz^2 \qquad (38)$$

where in empty space p_1, p_2, p_3 are such that $p_1 + p_2 + p_3 = p_1^2 + p_2^2 + p_3^2 = 1$. This metric is not conformally invariant so massless particles are created and also massive particles are created independently of their mass even in the ultrarelativistic case. Time derivatives of metric components are of the order of $1/t$, the curvature is of the order of $1/t^2$ and energy density of created particles is of the order of $1/t^4$. These quantities are diverging near singularity. We do not know how to handle this situation in the general case. It was easy when initially we had Minkowski space. Then one knows how to define the vacuum state and how to impose initial conditions. In the Kasner

space on the contrary the most important processes are going on near singularity. In order to avoid this difficulty Starobinsky assumed that up to some moment $t = t_0 > t_{Pl}$ space-time was given by Minkowski metric and at $t_0 > t_{Pl}$ it is smoothly matched with Kasner solution so initially we have again Minkowski space. In this case the rate of creation of particles is roughly $1/t_0^4$. Energy density of particles created at the initial moment is not enough to influence the background metric. It is small in comparison with the curvature of space-time. In the Kasner space-time there are always two directions of expansion and one direction of contraction but any element of volume is increasing so expansion dominates. Therefore energy density of particles created at initial moment is slowly decreasing with time. Momenta of particles moving along the axis of contraction are decreasing as $\frac{1}{t_0^4}\left(\frac{t}{t_0}\right)^{-p_1}\frac{1}{t} \sim t^{-1+|p_1|}$ and energy density of these particles is decreasing more slowly than curvature. Therefore though created particles at the first instant do not influence the background metric after some time they are becoming important.

In the cylindrical case of the Kasner metric there are two axes along which there is expansion and one along which there is contraction. The absolute value of scale on these axes has no meaning. At the moment $t = t_0$ we switch on the particle creation process (Lukash et al. (1976)). Initially created particles do not influence the background metric. After some time particles which are moving along the axis of contraction start to influence the background metric. Contraction is slowed down and finally space is forced to expand along this axis. At the same time expansion in other two directions is also slowed down and space starts to contract in these directions (see Fig. 2). In order to smoothly match Minkowski space with the Kasner metric at $t = t_0$ we can assume $a(t_0) = b(t_0) = c(t_0)$. After some time t_1 we will again have $a(t_1) = b(t_1) = c(t_1)$. At that moment of time momenta of particles moving in different directions are approximately equal. From the point of view of dynamics of space-time the moment $t = t_1$ is not exceptional and therefore expansion along the a axis and contraction along axes b and c will continue up to the moment when particles moving along axes b and c start to influence the metric. We see that along any axis periods of contraction follow periods of expansion and as a result of this the overall expansion becomes more isotropic (Lukash et al. (1976), Hu and Parker (1977)). The characteristic time in which anisotropy decreases is given by $t_1 \sim t_0\left(\frac{t_0}{t_{Pl}}\right)^{\alpha}$ where $\alpha \sim 1$. So if t_0 is close to t_{Pl} anisotropy is decreasing very rapidly. However due to streaming of particles along prefered axes small anisotropy

will persist for a long time. One can develope this theory much further but I do not think that it is necessary. This theory is based on one free parameter, namely t_o and I do not know of any theory which could predict the value of t_o.

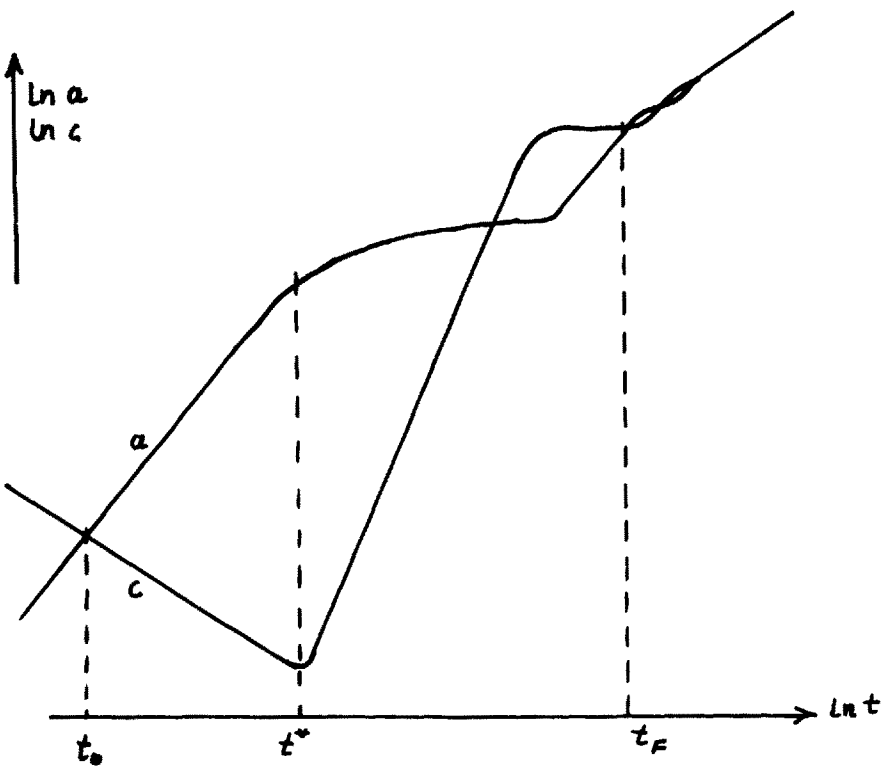

Fig. 2. Evolution of axisymmetric Kasner model. At $t = t_o$ particle creation process is turned on; at $t = t^*$ created particles start to influence the dynamics of expansion; t_F is the moment of isotropization; $A \sim t_o/t_{Pl} \sim 1$. Created particles are considered to be free particles.

I prefer another formulation. If $t_o \sim t_{Pl}$ then anisotropy decreases very fast so practically the expansion of universe instantly becomes isotropic. Therefore when we take into account the quantum

process of particle creation then it turns out that the anisotropic singularity is intrinsically inconsistent. The only sensible choice is to assume that the initial singularity is quasi isotropic (Lifshitz and Khalatnikov (1963)). Metric near the singularity could be written as

$$ds^2 = dt^2 - tg_{ij}(x)\, dx^i\, dx^j + t^2 k_{ij}(x)\, dx^i\, dx^j + \ldots \qquad (39)$$

and when g_{ij} is given one can calculate from Einstein equations all the higher corrections k_{ij} etc. Space-time described by the metric (39) is of course anisotropic but in comparison with curvature, which behaves like t^{-2} anisotropy is of the order of t^{-1} and it does not lead to creation of many particles.

There is still another situation which is interesting from the point of view of particle creation. I have in mind white holes. The initial big bang is not necessarily simultaneous. One can formulate initial conditions in such a way that expansion of some parts of universe are delayed. You have then a picture inverse to a black hole and therefore these objects, invented by Novikov (Novikov (1964)), are called white holes. Before explosion of a white hole there will be a singularity in the metric and it is possible to show that near this singularity, which is locally not distinguishable from cosmological singularity, particles will also be created. Even in the classical picture one can assume that some parts of matter will not start expanding before a given moment. This matter produces such a strong gravitational field that it is actually inside its own gravitational radius. As is well known, space-time inside gravitational radius is not static (it is also not stationary) and creation of particles is possible. In classical theory you can delay explosion of white hole for an arbitrary long time, but you can not do it if creation of particles by gravitational field is included.

In order to give complete picture of quantum particle creation process in gravitational field I should also mention that this process occurs also in gravitational field of a black hole. This was first investigated by Hawking (Hawking (1974), (1975)) who analyzed the behaviour of a scalar field in the vicinty of a collapsing star and showed that a distant observer will see a stream of particles with thermal spectrum, produced by gravitational field of a black hole. The rate of creation of particles depends on asymptotic parameters of black hole (mass, angular momentum, charge) and does not depend on the dynamics of collapse.

At the end I will discuss one technical point. We will consider

massless scalar field on a sandwich type space-time with Minkowski space for $t < t_1$ and $t > t_2$ ($t_2 > t_1$) and conformally flat space for $t_1 < t < t_2$. General state of the field can be decomposed into creation and annihilation operators and coefficients of this expansion are just the classical solutions of scalar wave equation. Classical solutions are important because even on classical level we can describe creation of particles. If initially we have some waves propagating from left to right and in the final state we have waves propagating from left to right and their amplitude is not changed and there are no waves propagating from right to left then we know that new particles were not created in the conformally flat region. However if we are interested in what is going on in perturbed but conformally flat region and we want to calculate the energy momentum tensor, then we have to deal with divergent quantities (DeWitt (1975)). In order to extract finite result we have to renormalize our theory. One way to do it is to introduce mass μ and recalculate all quantities for massive scalar field and only at the very end take a limit $\mu \to 0$. As was already mentioned before energy density of created particles is proportional to μ^2 and it will go to zero in the limit but we will notice that the trace of energy momentum tensor, after introduction of metric perturbations, in the limit $\mu \to 0$ is not zero but is proportional to $\ddot{h}$. This fact that the trace of renormalized energy momentum tensor of massless particles is not zero even in conformally flat space is called conformal anomaly (Deser, Duff, Isham (1976)). Conformal anomaly is always found in curved space and actually the trace of energy momentum tensor is proportional to the curvature of space-time. The conformal anomaly does not depend on renormalization procedure used and it is not connected with the field itself but rather with properties of vacuum polarization.

Let me make one more final comment. The number of particles created in the homogeneous case is

$$N = \int \ddot{h}^2 \, dt \tag{40}$$

and therefore

$$\frac{dN}{dt} = \ddot{h}^2 \tag{41}$$

but this number has a meaning only if one does not ask about the spectral distribution of created particles. It means that it is not possible at a given moment of time to distinguish uniquely the energy density of created particles from the energy density of vacuum polarization. When perturbation of space-time depends on space variables

we have

$$N = \int \left| h_{\vec{k},\omega} \right|^2 \theta(\omega - |\vec{k}|) \, d\omega \qquad (42)$$

where θ denotes the step function. Now we cannot transform this integral into a local quantity. Creation of particles is a non local process, it depends on the overall spatial picture and not only on derivatives of the metric at a given point.

References

Adler, S.L., Lieberman, J., Ng, Y.L., 1977, Ann. Phys. 106, 274.
Akulov, V.P., Volkov, D.V., 1973, Phys. Lett. B46, 109.
Bernard, G., Duncan, A., 1977, Ann. Phys. 107, 201.
Bjorken, J.D., Drell, S.D., 1965, Relativistic Quantum Fields, Mc-Graw Hill Book Company, New York.
Bogoliubov, N.N., Shirkov, D.V., 1959, Introduction to the Theory of Quantized Fields, Interscience Publishers, Inc., New York.
Davies, P.C.W., Fulling, S.A., Christensen, S.M., Bunch, T.S., 1977, Ann. Phys. 109, 108.
Deser, S., Duff, M.J., Isham, C.J., 1976, Nucl. Phys. B111, 45.
Deser, S., Zumino, B., 1976, Phys. Lett. 62B, 335.
DeWitt, B.S., 1975, Phys. Rep. 19C, 295.
Einstein, A., 1917, Phys. Zeits. 18, 121.
Ford, L.H., 1975, Phys. Rev. D11, 3370.
1976, Phys. Rev. D14, 3304.
Freedman, D.Z., van Nieuwenhuizen, P., Ferrara, S., 1976, Phys. Rev. D13, 3214.
Golfand, Yu.A., Likhtman, E.P., 1971, Lett. JETP, 13, 452.
Grishchuk, L.P., 1974, Sov. Phys. JETP 40, 409.
Hawking, S.W., 1970, Comm. Math. Phys. 18, 301.
1974, Nature 248, 30.
1975, Comm. Math. Phys. 43, 199.
Hawking, S.W., Penrose, R., 1970, Proc. Roy. Soc. Lond. A314, 529.
Hu, B.L., Parker, L., 1977, Anisotropy Damping Through Quantum Effects in the Early Universe, University of Wisconsin-Milwaukee preprint, UWM-4867-77-11.
Isham, C.J., Penrose, R., Sciama, D.W., Ed., 1975, Quantum Gravity, Clarendon Press, Oxford.
Lifshitz, E.M., Khalatnikov, I.M., 1963, Adv. in Phys. 12, 185.
Lukash, V.N., Novikov, I.D., Starobinsky, A.A., Zel'dovich, Ya.B., 1976, Nuovo Cimento B35, 293.
Lukash, V.N., Starobinsky, A.A., 1974, Sov. Phys. JETP 39, 742.
Novikov, I.D., 1964, Astr. Zh. 41, 1075.
Parker, L., 1977, in Asymptotic Structure of Space-Time, Ed. Esposito, F.P., Witten, L., Plenum Publ. Corp., New York.
1979, in Gravitation: Recent Developments, Ed. Levy, M., Deser, S., Plenum Publ. Corp., New York.
Schweber, S., 1961, An Introduction to Relativistic Quantum Field Theory, Row, Peterson and Comp., Evanston, Ill.
Sexl, R.U., Urbantke, H., 1967, Acta Phys. Austriaca 26, 339.
Starobinsky, A.A., Zel'dovich, Ya.B., 1972, Sov. Phys. JETP 34, 1159.
1979, Quantum Effects in Homogeneous Cosmological Models, in Bianchi Universes and Relativistic Cosmology, Ed.Ruffini,R.,Reidel, Holland.

Wald, R.M., 1977, Comm. Math. Phys. 54, 1.
Wess, J., Zumino, B., 1974, Nucl. Phys. B70, 39.
Woodhouse, N.M.J., 1976, Phys. Rev. Lett. 36, 999.
Zel´dovich, Ya.B., 1973, JETP 64, 58.

VISCOUS DISSIPATION AND EVOLUTION OF HOMOGENEOUS COSMOLOGICAL MODELS

Niccolò Caderni
Institute of Astronomy, Cambridge, U.K.

Ante mare et terras et quod tegit omnia caelum
unus erat toto naturae vultus in orbe,
quem dixere c h a o s : rudis indigestaque moles
nec quicquam nisi pondus iners congestaque eodem
non bene iunctarum discordia semina rerum.

Ovidius: Metamorphoseon
Book I vv. 5-9

1. Introduction and Viscosity Coefficients

In this report I shall consider the effects of viscous phenomena on the evolution of some anisotropic cosmological models. The existence of a highly dissipative epoch in the history of the universe was first claimed by Misner [1] in the hope of explaining the large scale regularity of the present universe without postulating strictly homogeneous and isotropic configurations since the beginning of the cosmological expansion. The conceptual difficulty (if not impossibility) of stating initial conditions on the initial singularity suggested that the status of the universe is not determined by very special a priori initial conditions, but is due to evolutionary processes closely linked to the physical properties of the cosmological matters itself. Such a philosophy, and the program generated by it was called "chaotic cosmology" [2,3]. It states the logical and physical inescapability of our universe which, born out from a primordial chaos, was able to smooth out any irregularity and to develop into its present configuration.

On the other hand, the practical difficulty of finding physical processes actually effective to drive the cosmological evolution from any initial condition up to the present status generated a competing set of theories which we shall call "quiescent cosmologies" [4]. In this context several intriguing theories are brought in support of the idea that the universe was regular since the remote past.

The major source of cosmological information is the microwave background radiation. The investigation of its spectrum [5] shows that the universe is expanding quite isotropically at least since the time in which it became transparent to the radiation and that the ratio of the background photon density to the baryon density in the universe is of order $\sim 10^8$. A simple calculation shows that this number can be interpreted as the radiation entropy per particle in the expanding universe [6].

The explanation of the isotropy and of the entropy content of the universe therefore becomes the battelfield of the competing cosmological theories. Without any pretext of completeness and generality, having only the purpose of inserting the following work into a slightly more general context, I tried to indicate in fig.(1)how modern cosmology poses itself in regard to the problem of initial conditions. It is of course impossible to translate the greek philosophy into modern scientific thoughts and methodology. Nevertheless I call "Epicurean" those models where the primordial chaos, owing to successive μεταμόρφωσις is now reduced to our almost regular universe [7]. On the other hand, models where a strict symmetry exists since the beginning may properly be called "Aristotelian" [8]. I will quote some ideas supporting each view and the reader is sent to the current literature for direct explanations.

On the Aristotelian side one can first find an hypothesis mainly due to Penrose [9] in which the isotropy of the space is associated with a gravitational entropy expressed by the Weyl tensor. Barrow and Matzner's [10] idea is that the number $\sim 10^8$ for the photon-baryon ratio instead of being large is actually too small if the universe was initially anisotropic. Owing to dissipative processes too much entropy would be generated in a chaotic universe. Finally, the Anthropic Principle [11] states, roughly speaking, that the isotropy of the universe is a mere consequence of our existence, as well as the fact that the universe is expanding at just the critical rate. Among the infinity of possible universes only a quite regular one would allow life to be developed in it.

On the Epicurean side several degrees of chaos may be found [12]. The first distinction is between homogeneous and inhomogeneous models. We know that inhomogeneities do exist, although, due to their relative mathematical simplicity, homogeneous models are more often investigated in the current literature. Nevertheless anisotropic but homogeneous models allow one to investigate several physical processes which do not occur in Friedman universes. In models in which a perfect fluid

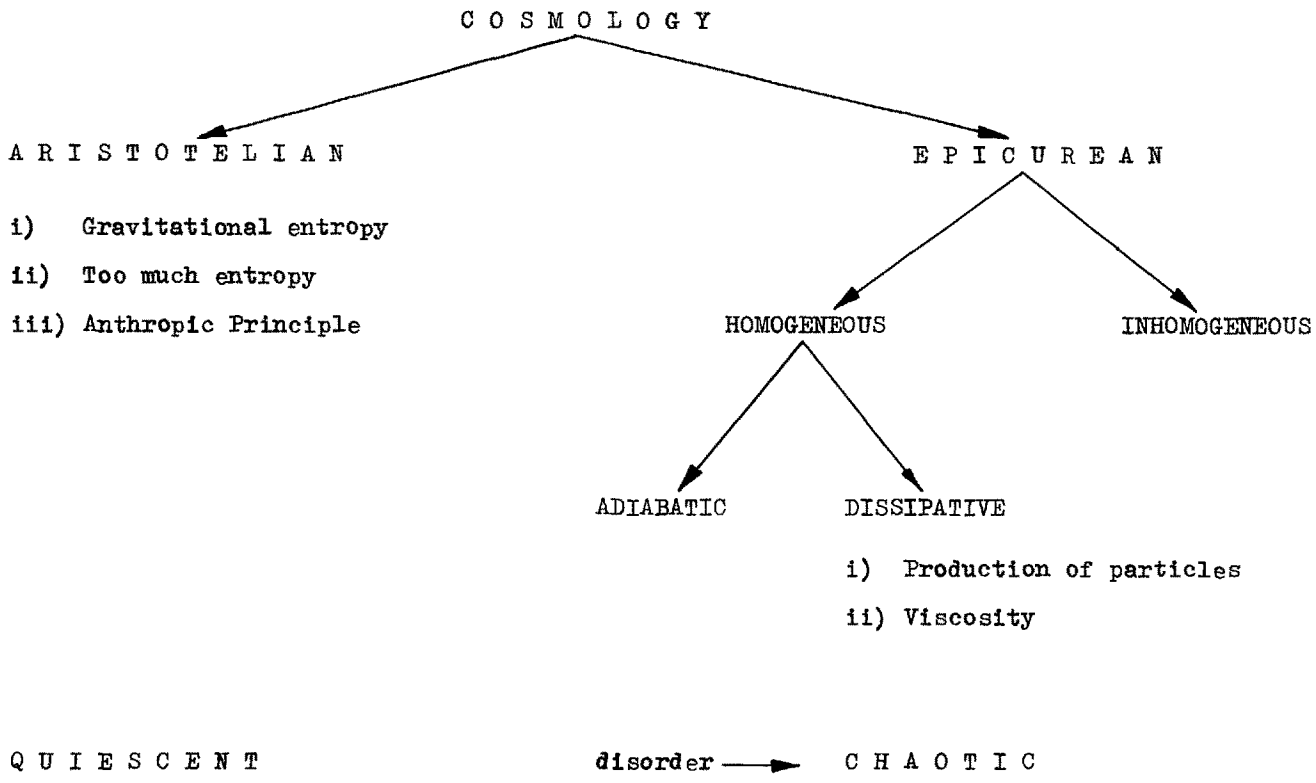

Fig. 1. Modern cosmology and the problem of initial conditions

is assumed to be the source term in the Einstein equation and the evolution toward isotropic configurations is merely adiabatic [13,14], some essential physical properties of the cosmological matter may be ignored: as a matter of fact the initial anisotropy may be strongly damped by various kinds of dissipation, leading to a large increase of the radiation entropy. Two kinds of dissipative interactions can take place between particles and fields in the first moments of the Universe's life. The first one is a quantum effect, namely the creation of pairs of particles in time-dependent gravitational fields [15,16]. Pair production takes place at very early times (i.e. close to the Planck time $t_p \sim 10^{-43}$ sec) and is supposed to be very effective in spite of the short duration.

The other dissipative process is viscosity [1,17-28]. If the mean free paths of particles in the cosmological fluid are long enough, viscous forces are able to damp the expansion anisotropy quite effectively. Such a situation arises mainly in the lepton era of the Universe, when the temperature was $6 \cdot 10^9$ K $\lesssim T \lesssim 1.5 \cdot 10^{12}$ K, because of the weak coupling between neutrinos and electron-positron pairs. A previous dissipative era might take place at much higher temperatures ($T \sim 10^{20}$ K) [29] when gravitons are assumed to be in thermal equilibrium with matter. However the existence of such an epoch is problematic because of a possible ultimate temperature $T \simeq 10^{12}$ K [30]. Quintessentially epicurean, inhomogeneous models have not been extensively studied, although a large portion of future cosmology may reside therein [12].

One should notice that the debate between epicurean and aristotelian cosmologies is actually going on in other, very important, astrophysical subjects, in which we are not here involved. In particular it is not true that the present structure of the universe would arise from an arbitrary spectrum of the primordial fluctuations. On the contrary, only special assumptions about the shape and the amplitude of such fluctuations do allow protostructures to be formed, according to the present galaxy formation theory.

In the following I shall restrict myself to the investigation of viscous phenomena in the lepton era, giving a slightly literal review of a series of papers by Roberto Fabbri and the author, here quoted as [31-36]. The rest of this chapter is devoted to the calculation of viscosity coefficients for the neutrino-neutrino and the electron-neutrino plasma. Then, in chapter 2, the main features of the dissipative process, are displayed in the simple case of the flat, Bianchi type I space. When we will try to extend our conclusions to

more general curved models, a peculiar coupling between anisotropy and curvature featured by these models will rebuff the former optimism: Chapter 3 deals with the evolution of Bianchi type IX and type VIII models. Lastly, a simple case of an open model (type V) is studied in Chapter 4 together with the final non-conclusions.

Though obviously an over-simplification it makes sense to approximately describe the cosmological fluid during the lepton era as a non-degenerate mixture of electrons and (electron) neutrinos. Accurate expressions for the various viscosity coefficients may be found by the method of relativistic quantum kinetic theory and by the knowledge of the weak interaction cross-sections below 150 Mev. A fully covariant kinetic gas theory has been developed during last years by a number of scientists in Amsterdam [37-38]. Such a theory, which applies to very hot dilute gas, is based upon a generalization of the Enskog approximation and provides explicit expressions for the coefficients characterizing transport in binary gas mixture. These expressions have the form of successive approximations and are valid for all temperatures. More recently such a theory has been modified, taking into account quantum processes [39-40].

The energy momentum tensor for a viscous fluid may be written as

$$\begin{aligned} T_{\alpha\beta} = & \left[\varepsilon + p - \left(\eta_v - \frac{2}{3}\eta_s\right) u^{\mu}{}_{;\mu}\right] u_\alpha u_\beta \\ & + \left[p - \left(\eta_v - \frac{2}{3}\eta_s\right) u^{\mu}{}_{;\mu}\right] g_{\alpha\beta} \\ & - \eta_s \left(u_{\alpha;\beta} + u_{\beta;\alpha} + u_\beta u^\lambda u_{\alpha;\lambda} + u_\alpha u^\lambda u_{\beta;\lambda}\right) \end{aligned} \qquad (1-1)$$

where u_α is the hydrodynamic four velocity, normalized as $u^\mu u_\mu = -c^2$, η_s and η_v are the shear and the volume viscosity respectively. Here and henceforth, latin indices assume the values 1, 2, 3, whereas greek indices assume the values 0, 1, 2, 3. According to relativistic quantum kinetic theory, given the interaction between the constituent particles of the system, the viscosity coefficients may be calculated in successive approximation. They can be expressed in terms of functions F_{ab} defined as

$$F_{ab}(z) \equiv \int y^{-a} \left(y^2 - 1\right)^b K_n(zy)\, dy$$

$$n = \frac{5}{2} + \frac{1}{2}(-1)^a$$

where $K_n(x)$ is the modified Bessel function of the second kind of or-

der n and $z \equiv mc^2/kT$ with m the electron mass and k the Boltzmann constant. Taking the composition of the cosmological (e,ν) mixture such that the electron number density is twice that of neutrinos (see however ref.[24]), we may write the first non-vanishing approximation to the viscosity coefficients in the form:

$$\eta_s = 15360\ \pi\ G^{-2}\ m^{-1}\ \hbar^4\ c\ z^3\ K_2(z)/S(z)$$

$$\eta_v = 32\ \pi\ G^{-2}\ m^{-1}\ \hbar^4\ c\ z^3\ (4 - 3\gamma)^2\ K_2(z)/V(z)$$

where G is the weak coupling constant and $\gamma = c_p/c_v$. The quantities S(z) and V(z) are collision integrals depending on the details of the particle interaction. Thus, the value of the viscosity coefficients varies with the parameters of the weak interaction theory. In fact, on the basis of the Weinberg-Salam model we have:

$$S(z) = 942080\ z^2\ K_2(z) + z^9\left[(1+c)\left(\tfrac{1}{4}\ z^2\ F_{68} + 2z\ F_{78} + \tfrac{16}{3}\ F_{88}\right)\right.$$

$$- c\left(\tfrac{1}{8}\ z^2\ F_{88} + \tfrac{17}{12}\ z\ F_{98} + \tfrac{7}{2}\ F_{10,8}\right) + c^2\left(\tfrac{1}{16}\ z^2\ F_{68} + \tfrac{7}{24}\ z\ F_{78}\right.$$

$$\left.\left. + \tfrac{11}{12}\ F_{88}\right) + c^2\left(\tfrac{3}{160}\ z^2\ F_{10,10} + \tfrac{11}{40}\ z\ F_{11,10} + \tfrac{13}{20}\ F_{12,10}\right)\right],$$

$$V(z) = z^8\left[\left(\tfrac{1}{2} + \tfrac{1}{2}\ c + \tfrac{1}{12}\ c^2\right)\ F_{56} - \tfrac{1}{3}\ c\ F_{76} + \tfrac{1}{16}\ c^2\ F_{98}\right]$$

where $c \equiv 4\ \sin^2\theta_w$, with θ_w the Weinberg mixing angle. Different values for the both viscosity coefficients are shown on fig. (2), where the viscosities according to the charged current V-A theory are also indicated. It is seen that the value of the W-S viscosities exceeds the V-A value, but the Weinberg angle has the effect of reducing such an enhancement.

In the following, in order to integrate the Einstein equations for the cosmological models numerically, polynomial approximations for the viscosity coefficients are employed. For the W-S viscosities, in the case $\sin^2\theta_w = 0.35$ we will use the asymptotic expressions:

$$\eta_s = \frac{15360\,\pi^2}{2570250} \cdot \frac{\hbar^4 c}{G^2 m}\ z\ (1 - 0.0704\ z^2 + 0.0095\ z^4) \qquad (1\text{-}2a)$$

$$\eta_v = \frac{32\,\pi}{184861} \cdot \frac{\hbar^4 c}{G^2 m}\ z^5\ (1 + 0.7007\ z^2 - 0.8368\ z^4) \qquad (1\text{-}2b)$$

which in the range $z^{-1} > 1$ approximate the actual coefficients to within 1.5 % and a factor ~ 3 respectively. We remark that, throughout

this range $\eta_s/\eta_v > 10^2$ (see Fig. 2), so that the inaccuracy of the approximation to the bulk viscosity (which, for $z \leq 0.1$ is already less than 10%) is relatively unimportant.

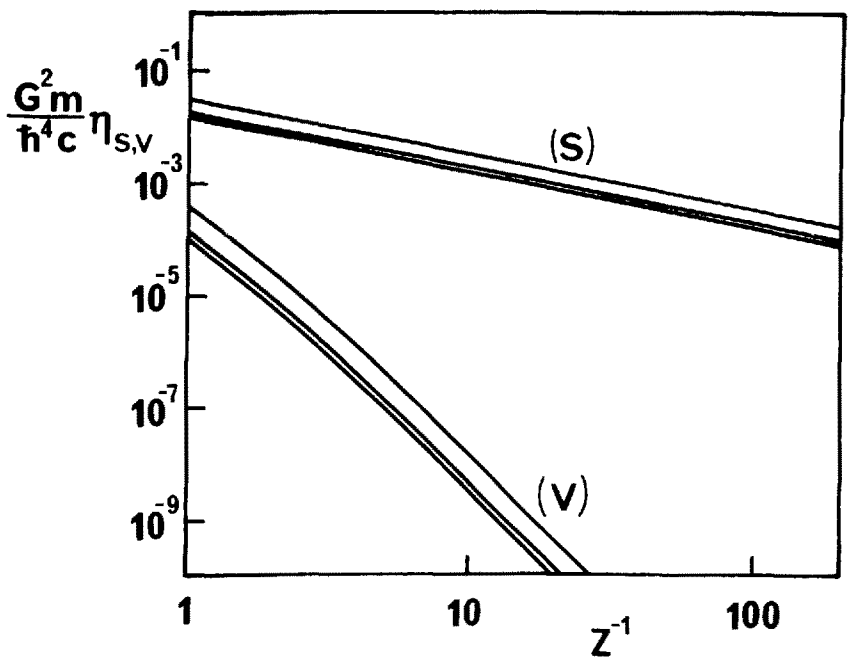

Fig. 2. Dimensionless coefficients of shear viscosity (S) and bulk viscosity (V) as functions of the reduced temperature $z^{-1} \equiv kT/mc$. In both sets of three curves the lowermost curve is based on V-A theory, while the other two are obtained in the context of the Weinberg-Salam model and represent the cases $\sin^2\theta = 0$ (upper curve) and $\sin^2\theta = 0.35$ (middle curve). (Ref. [33])

2. Neutrino Viscosity in Bianchi type-I Universes

This section will deal with the evolution of Bianchi type-I cosmological models under the influence of lepton viscosity. In geometrical units ($c = 8\pi G = 1$) the metric of such spaces can be written as

$$ds^2 = -dt^2 + \sum_{k=1}^{3} \left[R_k(t)\, d\mathscr{x}^k \right]^2 \qquad (2\text{-}1)$$

where $R_k(t)$ denotes the cosmic scale factor associated with the k-th

principal direction. Therefore the directional Hubble parameter H_k can be defined

$$H_k = \frac{d}{dt} (\ln R_k)$$

The Einstein equations for a space with the metric (2-1), filled with a fluid of shear viscosity η_s and volume viscosity η_v, can be reduced to a system of two coupled ordinary differential equations. They take the form:

$$\frac{dH}{dt} = \frac{1}{2} (\mathcal{E} - p) - 3H^2 + \frac{3}{2} \eta_v H \qquad (2\text{-}2a)$$

$$\frac{d\mathcal{E}}{dt} = -3H (\mathcal{E} + p) + 9 \eta_v H^2 + 4 \eta_s (3H^2 - \mathcal{E}) \qquad (2\text{-}2b)$$

where H is the average Hubble parameter $H = (H_1 + H_2 + H_3)/3$, $\mathcal{E}$ is the energy density and p the pressure of the cosmological fluid. Equations (2-2) were first investigated by Belinsky and Khalatnikov [25] in the case of hypothetical dependence of the viscosity coefficients on the energy density, using the technique of dynamical systems. Here we integrate the Einstein equations numerically. We assume that the cosmological matter can be represented at each epoch by a fluid in thermal equilibrium, and we adopt the viscous approximation using the coefficients found in Chapter 1. In the lepton era the main contribution to the energy density comes from electron-positron pairs, neutrinos and photons. The contribution of matter (ionized hydrogen) is totally negligible, therefore the equation of state is $p = \mathcal{E}/3$. All kinds of particles are in thermal equilibrium, and it is possible to define a unique temperature T, such that $\mathcal{E} \propto T^4$.

The viscosity can strongly modify the dynamics of the evolution of a cosmological model. In eq. (2-2b) three terms contribute to the rate of change of the energy density $\mathcal{E}$: the first one describes the adiabatic cooling of an expanding gas, whereas the others give the heating due to viscous forces. The bulk viscosity term provides an additional pressure $P_v = -3 \eta_v H$ and it is already present in isotropic models as the only kind of viscosity allowed therein. Its influence on Friedman universes has been investigated, in the framework of the present investigations, in ref. [31] . On the other hand, the shear viscosity term depends on the geometrical configuration and has a much greater bearing on the cosmological evolution, its contribution is dominating in anisotropic Bianchi spaces. As a matter of fact the expansion anisotropy A defined as

$$A = \frac{1}{3} \sum_{k=1}^{3} \left(\frac{H_k - H}{H} \right)^2 \qquad (2\text{-}3)$$

in type-I spaces takes the form

$$A = 2 - \frac{2\varepsilon}{3H^2}$$

so that the amplification factor for the shear viscosity ($3H^2 - \varepsilon$) is nothing else but the anisotropy energy density ε_a

$$\varepsilon_a = \frac{3}{2} AH^2 \qquad (2\text{-}4)$$

So, the more anisotropic the space becomes, the more important the viscous heating term is. Therefore, one is forced to think about the possibility of having solutions in which the space is highly anisotropic, i.e. $\varepsilon \ll 3H^2$, the positive η_s-dependent term is larger in magnitude than the negative one and the viscous heating prevails over the expansion cooling. Therefore, the use of the viscous fluid scheme allows us to define two classes of solutions. In the former, which we call class A, we include solutions in which the lepton era, following a previous hadron era, starts at $T \simeq 1.5 \cdot 10^{12}$ K. For all of these world models the temperature decreases monotonically with time. In the latter class (class B) we include solutions in which the lepton era starts at $T \simeq 6.10^9$ K. Here the temperature increases very rapidly for a short time, then drops down.

However, the use of viscous approximation in cosmological problems has been questioned by several authors [17,18]. The most restrictive criterium for its validity is provided by the Stewart's theorem [18]. This theorem poses a limit in the rate at which the energy density T^{oo} (radiation + matter) can be increased by the anisotropy of space; in other words it restricts the effectiveness of the viscous mechanism. The Stewart inequality reads:

$$\left| \frac{d(\ln T^{oo} R^4)}{d(\ln R^2)} \right| \leqslant 1 \qquad (2\text{-}5)$$

where R is the average scale factor. As far as class-A models are concerned the above condition is fulfilled throughout the lepton era. The collision time t_c becomes larger than the hydrodynamic time H^{-1} just at $T \sim 10^{10}$ K, but it has been proved [24] that the viscous heating does not cease at $t_c H \sim 1$. On the other hand, Stewart's upper limit is apparently in contradiction with the existence of regions with $d\varepsilon/dt > 0$, found in all class-B solutions. However, a consist-

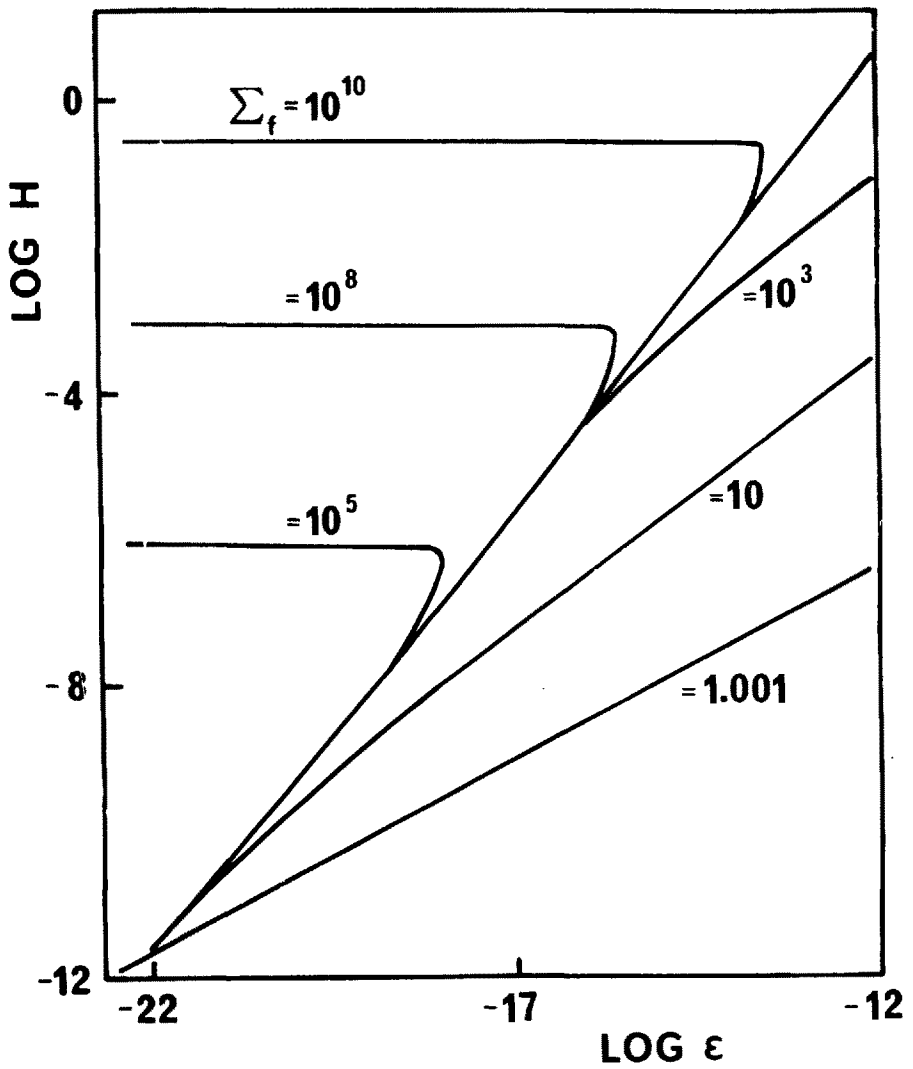

Fig. 3. The Hubble parameter as a function of the energy density for the lepton era of a few Bianchi type I models. Each curve is labelled by the corresponding final entropy ratio Σ_f . The value $\Sigma_f = 1.001$ corresponding to the Friedman solution. (See ref. [32]) .

ent treatment of the dissipative mechanism [12,24] shows that, when $t_c H \gg 1$, the anisotropy energy density can be stored in a "potential" form and in this form it already resides in the T^{oo} so that one can write $T^{oo} = \mathcal{E}_{iso} + \mathcal{E}_{aniso}$, with the possibility of having $\mathcal{E}_{aniso} \gg \mathcal{E}_{iso}$. If we use the fluid approximation only $\mathcal{E}_{iso}$ is ascribed to T^{oo}, whereas $\mathcal{E}_{aniso}$ is formally included in the anisotropy energy density. Therefore eq. (2-2b) refers to the rate of change of the isotropic energy density only. In this way the Stewart's theorem may be circumvented: we must identify the quantity $\mathcal{E}$ appearing in eqs (2-2) with the isotropic energy density $\mathcal{E}_{iso}$. While $\mathcal{E}_{iso}$ increases sharply in the class B models, the total T^{oo} lie in the range imposed by the inequality (2-5). Moreover, we shall also identify in the following the temperature with the isotropic temperature $T \propto \mathcal{E}_{iso}^{1/4}$. Nevertheless, in class B solutions these may lead to important errors in the first steps of the integration. As a matter of fact, the important feature displayed by these unconventional models is that the universe springs for a while into a collisionless regime. Afterwards, when the temperature is $T \sim 10^{10}$ K, the neutrinos become collision-dominated.

Solutions of both classes are clearly displayed in the $(H, \mathcal{E})$ plane which represents the phase space of the dynamical system (2-2). In fig.(3) we report few models of both classes. The separatrix of the above solutions is the curve $H = M\, \mathcal{E}^{1.27}$ with $M = 4.1 \cdot 10^{16}$. A local maximum of $\mathcal{E}$, considered as a function of $H \sim t^{-1}$, appears very clearly in each class B models: it corresponds to the maximum (isotropic) temperature, which viscous heating is able to produce before being overcome by the expansion cooling.

In geometrical units the initial values of $\mathcal{E}$ for both classes are $\mathcal{E} = 3.91 \cdot 10^{-13}$ cm^{-2} for class A, and $\mathcal{E} = 7.04 \cdot 10^{-23}$ cm^{-2} for class B. The initial values of the Hubble parameter H_{in} for class A lie between the isotropic solution $\mathcal{E} = 3H^2$ and the separatrix. For class B, H_{in} is limited from below by the condition $\mathcal{E}_{in} > 0$. Then we allow H to increase till the anisotropy is so high to bring the isotropic temperature up to $T \simeq 1.5 \cdot 10^{12}$ K. A larger anisotropy would drive the universe in a hadron era that we are not able to treat, but where, due to the short mean free paths of the strong interactions, viscous phenomena are not expected to play an important role. The range of initial conditions for H is then

$$\begin{aligned} &3.4 \cdot 10^{-7}\ \text{cm}^{-1} \lesssim H_{in} \lesssim 15\ \text{cm}^{-1} && \text{Class A} \\ &10^{-11}\ \text{cm}^{-1} \lesssim H_{in} \lesssim 10\ \text{cm}^{-1} && \text{Class B} \end{aligned} \qquad (2\text{-}6)$$

Fig.(3) also shows that for any initial condition in class B and for class A models with $H > 10^{-4}$ the universe is represented by the same point in the phase space at the end of the lepton era: the anisotropy is reduced to $A \sim 0.82$ and the subsequent evolution is practically the same for all the cases.

Integrating equations (2-2) for the plasma era one should introduce as an initial condition a small contribution $\mathcal{E}_m$ to the energy density, representing the rest mass density of matter, which will become important in later stages. Moreover the equation of state is now $p = \frac{\mathcal{E}_{rad}}{3}$, where $\mathcal{E}_{rad}$ is the radiation energy density only. According to our investigation, viscous phenomena, here due to the photon-electron interaction [31], are not able to influence the cosmological evolution for $T < 6 \cdot 10^9$ K. All dynamical variables evolve in the same manner as in adiabatic model up to 10^{-4}. By choosing a suitable value for the matter density contribution $\mathcal{E}_m$ we get the value $H = 50$ km sec^{-1} Mpc^{-1} when the background temperature is $T = 2.7$ K. This value of $\mathcal{E}_m$, which is the same for all the cases, corresponds to a baryon number density $N \sim 10^{23}$ cm^{-3} at $T = 6 \cdot 10^9$ K and just the critical density $N \sim 10^{-6}$ cm^{-3} at the present time. The anisotropy of space decreases further according to the adiabatic expansion law $A H^2 R^6 = \text{const}$, so one can calculate the expected quadrupolar anisotropy of the background radiation. We found

$$\frac{\Delta T}{T} = (3.8 \pm 0.3)\ 10^{-7}$$

This result, obtained for a large set of initial anisotropies, shows how effective the neutrino viscous damping is. For comparison, an adiabatic universe with $2 - A = 10^{-17}$ at $T = 10^{12}$ K would exhibit an anisotropy $\frac{\Delta T}{T} \sim 0.89$ at the present time. The best accuracy in the measurement of the quadrupole anisotropy of the cosmic background is now $\sim 10^{-3}$, thus anisotropies as small as 10^{-7} are completely unobservable.

The large damping of the anisotropy which takes place during the few seconds of the lepton era is accompanied by a huge enhancement in the radiation content of the universe. The entropy production may be studied by looking at the ratio of the radiation entropy at some time t to the initial radiation entropy:

$$\Sigma(t) = \left(\frac{R(t)\ T(t)}{R_{in}\ T_{in}} \right)^3 \qquad (2\text{-}7)$$

Because of the strict connection between the viscous dissipation and the damping of anisotropy we are able to find simple relations between

the geometrical quantities and the physical properties of the cosmological matter. This is particularly true for class-B models where the radiation entropy enhancement via viscous dissipation is at each stage governed by the law

$$\Sigma = 2\frac{H_{in}}{H}\left(\frac{\mathcal{E}}{\mathcal{E}_{in}}\right)^{1/2} = 2\left(\frac{2-A}{2-A_i}\right)^{1/2} \qquad (2\text{-}8)$$

We remember that class B models are defined as $\mathcal{E}_{in} = \mathcal{E}_f$, where the subscript f means end of the lepton era, therefore we found the following expression for the final entropy ratio

$$\Sigma_f = 2\frac{H_{in}}{H_f} \qquad (2\text{-}9)$$

Thus we arrive at the following remarkable result: due to neutrino viscosity the radiation entropy increases by a factor which depends only on the initial anisotropy. In class B models H_{in} is limited from above by the condition $H_{in} \simeq 10\ cm^{-1}$ so, according to our calculations, the maximum entropy production via neutrino viscosity in the lepton era is $\Sigma_f \simeq 2 \cdot 10^{12}$. In class A models the photon number density increases but not so much and a simple power-law dependence between entropy and anisotropy is only available for sufficiently large initial anisotropies (i.e. $H_{in} \gg 10^{-4}$), so

$$\Sigma_f = 10^{-17}\ \frac{\mathcal{E}_{in}}{\mathcal{E}_f}\ \frac{H_{in}}{H_f} \qquad (2\text{-}10)$$

The maximum entropy ratio for these models corresponds to the separatrix, where $\Sigma_f \cong 10^5$. In the isotropic solution, where only the bulk viscosity is present, we found an entropy production $\Sigma_f = 1.0016$.

Once we realized how important for the cosmological evolution viscous phenomena can be, we may ask how our results depend on the magnitude of viscosity coefficients. In view of the fact that viscous approximation can break down at some stage, it is important to see what happens when viscosity coefficients are much less than those obtained in Chapter 1. Let us consider the dependence of the final entropy ratio Σ_f and the final anisotropy A_f on the viscosity coefficients by setting

$$\tilde{\eta}_s = K_s\,\eta_s \qquad\qquad \tilde{\eta}_v = K_v\,\eta_v$$

where K_s and K_v are constant. The results are reported in fig. (4) where $K = K_s = K_v$ lie in the range $10^{-2} \lesssim K \lesssim 10$. Three main features

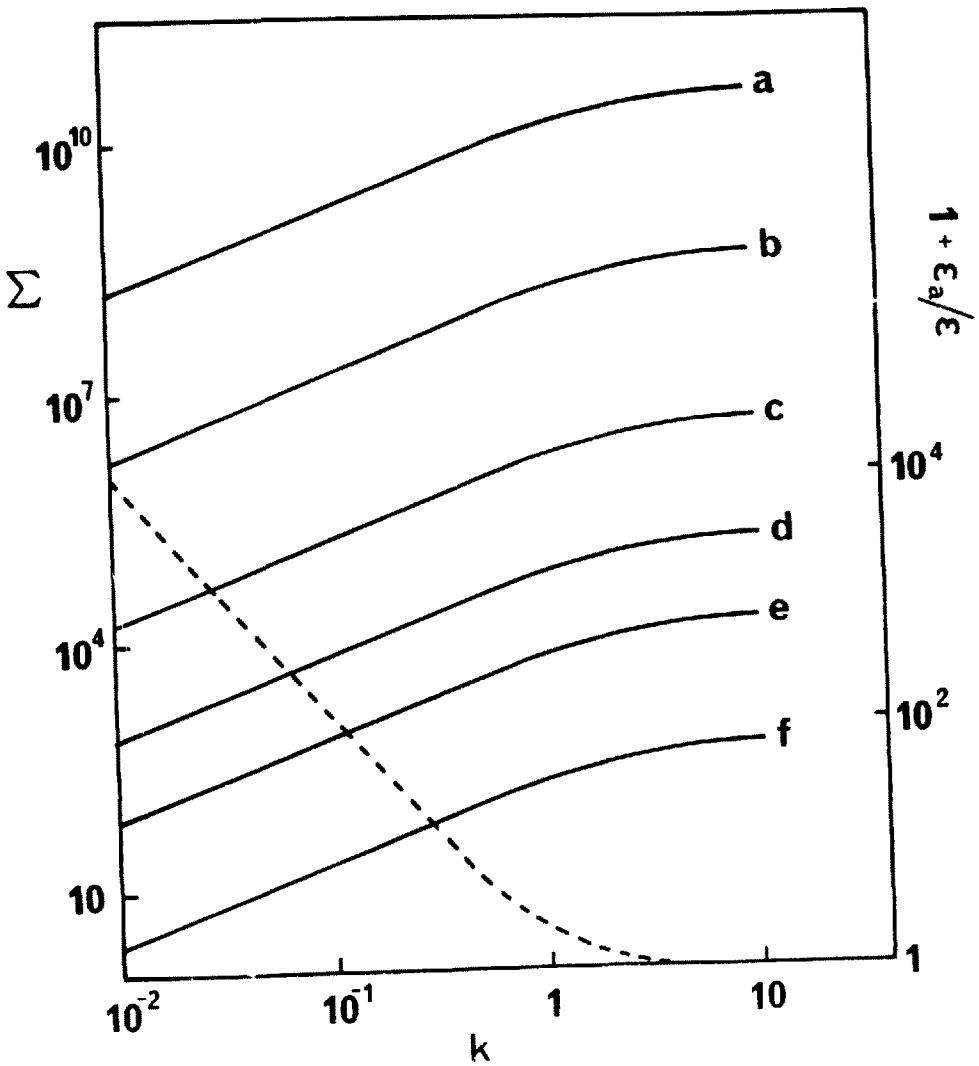

Fig. 4. The final entropy ratio Σ_f (solid lines) and the final anisotropy $(\varepsilon_a/\varepsilon)_f$ (dashed) as functions of $K = \eta_s/\eta_s^{w-s}$. The curves for Σ refer to different values of the initial anisotropy $I = (\varepsilon_a/\varepsilon)_{in}$ namely (A or B denotes the class defined in the text): (a) B, $I = 3\times10^{20}$ (b) B, $I = 3\times10^{16}$, (c) B, $I = 3\times10^{12}$, (d) A, $I = 3\times10^{14}$, (e) A, $I = 3\times10^{11}$, (f) A, $I = 3\times10^{9}$. (See ref. [33]).

are apparent: the final entropy ratio (considered for several values of the initial anisotropy) is a linear function of K in the $K \lesssim 1$ region. This fact proves that, although the quantitative results depend on the magnitude of η_s, we would still find a large production of entropy even of the shear viscosity were substantially smaller than η_s^{w-s}, provided the universe was suitably anisotropic at the beginning of the lepton era. In particular strong dissipation takes place even if the neutrino concentration is much smaller than the electron concentration [24]. On the other hand, for $K \gtrsim 1$, a change of regime occurs and Σ depends on K weakly, suggesting the existence of a saturation limit. Such a limit prevents the production of entropy in the finite range of temperature of the lepton era from being infinite, even if we let A arbitrarily large at the beginning.

On dashed line we give the final anisotropy parameter which does not depend on the initial anisotropy. Since a cosmological model may be considered quasi-isotropic when $\mathcal{E}_a/\mathcal{E} \leqslant \frac{1}{2}$, the figure shows that the isotropization via neutrino viscosity is quite effective for $K \gtrsim 1$. Conversely, for smaller values of the shear viscosity, and especially for $K \lesssim 0.1$, the cosmological expansion would remain highly anisotropic if initially $\mathcal{E}_a/\mathcal{E} \gg 1$. Therefore the actual strength of the weak interactions is sufficient to permit the isotropization of a wide set of world models.

3. Neutrino Viscosity in Bianchi type-IX and type-VIII Universes

Computation of the viscosity coefficients for the lepton gas, an idealized constituent of the Universe in the first three seconds, allowed us to investigate in Chapter 2 the evolution of a wide range of cosmological models of the Bianchi type-I and to observe how the dissipative mechanism was able to damp out any primeval anisotropy and to substantially enhance the radiation content of the Universe. Therefore these conclusions support the chaotic cosmology program. However Bianchi type-I model does not possess a sufficient number of internal degrees of freedom [14,41] to allow conclusions drawn therein to be considered general. In other words Bianchi-I model is a zero-measure subset in the set of all homogeneous models, so that one is compelled to check whether the hypothesis that dissipative processes could smooth out any primeval irregularity works in more general situations. The most generic sets of homogeneous world models are Bianchi types VI, VII, VIII and IX: in this chapter we study the evolution of diagonal type-VIII and type-IX spaces. Type-IX models are generali-

zations of the k = +1 Friedman model and their evolution has been studied in ref. [41,12,13,14] ; type-VIII (semi-closed topology) does not contain strictly isotropic subspaces, but there are solutions quite plausible from the point of view of observational data.

The integration of the Einstein equations for both models clearly shows that, although neutrino viscosity does damp primeval anisotropy quite effectively, "new" anisotropy is pumped on by the curvature of space when the universe departs from a pure Kasner like behaviour. The existence of such a coupling between curvature and anisotropy prevents neutrino viscosity from being an effective agent of isotropization for closed and semi-closed topologies. If the universe was highly anisotropic at the beginning, strong upper limits on the initial curvature are posed by the observed isotropy of the background radiation. Moreover, the negative result about isotropization in curved models is independent on the dissipative process and it is a general property of homogeneous cosmologies, so that the status of chaotic cosmology turns out to be weakened.

In the diagonal Bianchi IX and VIII models the metric may be written in an orthonormal tetrad

$$ds^2 = - dt^2 + \sum_{i=1}^{3} (\omega^i)^2 \qquad (3\text{-}1a)$$

$$\omega^1 = R_i(t)\, \Omega^{\,i}, \quad \omega^0 = dt \qquad (3\text{-}1b)$$

where $\Omega^{\,i}$ are time-independent differential 1-forms. The exterior derivatives of the differential forms (3-1b) are related to the group structure of the three-space and can be written as

$$\begin{aligned} d\omega^i &= - \tfrac{1}{2}\, \mathcal{C}^i_{\alpha\beta}\, \omega^\alpha \wedge \omega^\beta \\ d\Omega^i &= - \tfrac{1}{2}\, c^i_{kl}\, \Omega^{\,k} \wedge \Omega^{\,l} \end{aligned} \qquad (3\text{-}2)$$

where the c^i_{kl} are the canonical structure constants and $\wedge$ denotes the exterior product. In our case:

$$c^l_{ik} = - \mathcal{E}_{ikl}\, c_l \qquad (3\text{-}3a)$$

where $\mathcal{E}_{ikl}$ is the skew pseudotensor and

$$\begin{aligned} c_1 &= 1 \qquad && \text{(type IX)} \\ c_1 &= c_2 = - c_3 = 1 \qquad && \text{(type VIII)} \end{aligned} \qquad (3\text{-}3b)$$

Using eqs. (3-1) and (3-3) one can derive the commutation coefficients $\mathcal{C}^{\gamma}_{\alpha\beta}$ for the unit vector of the orthonormal frame (3-1a),

$$\mathcal{C}^{k}_{ok} = - H_k$$
$$\mathcal{C}^{l}_{ik} = - \varepsilon_{ikl} \frac{c_l R_l}{R_i R_k} \tag{3-4}$$

One can now calculate the affine connection 1-form $\omega_{\mu\nu}$ and the curvature 2-form $\Re_{\mu\nu}$ with the standard formulas

$$\omega_{\mu\nu} = \frac{1}{2} (\mathcal{C}_{\mu\nu\alpha} + \mathcal{C}_{\alpha\mu\nu} - \mathcal{C}_{\nu\alpha\mu}) \omega^{\alpha} \tag{3-5a}$$

$$\Re_{\mu\nu} = \omega_{\mu}{}^{\alpha} \wedge \omega_{\alpha\nu} + d\,\omega_{\mu\nu} \tag{3-5b}$$

and the curvature tensor $R^{\mu}{}_{\nu\alpha\beta}$

$$\Re_{\mu}{}^{\nu} = R_{\mu}{}^{\nu}{}_{\alpha\beta}\, \omega^{\alpha} \wedge \omega^{\beta}$$

then calculate all the components of the Einstein tensor $G_{\alpha}{}^{\beta}$ which turn out to be diagonal. In order to construct the Einstein equations one needs also the stress-energy tensor $T_{\alpha\beta}$ for a viscous fluid as given in (1-1). Using the affine connections (3-5) we find for a fluid at rest in that frame

$$T_{oo} = \mathcal{E}$$
$$T_{kk} = p + (2\eta_s - 3\eta_v) H - 2\eta_s H_k \tag{3-6}$$

Unlike the type-I case, the non isotropy of the curvature tensor does not allow in the present case a unique differential equation for the average Hubble parameter to be written down; therefore the Einstein equations take the form: (a detailed derivation is given in ref. [35])

$$\frac{dH_k}{dt} = - 3HH_k + \frac{1}{4} \sum_{i \neq l \neq k} \left[\left(\frac{c_l R_l}{R_i R_k} - \frac{c_i R_i}{R_l R_k} \right)^2 - \left(\frac{R_k}{R_i R_l} \right)^2 \right]$$
$$+ \frac{1}{2} (\mathcal{E} - p) + \frac{3}{2} \eta_v H + 2 \eta_s (H - H_k) \tag{3-7a}$$

$$\frac{1}{2} \sum_{i \neq j} H_i H_j + \frac{1}{2} R^{*} = \mathcal{E} \tag{3-7b}$$

The space curvature R^{*} is nothing else but the trace of spatial components of the Riemann tensor, once one formally sets $H_k = 0$, and it

takes the form

$$R^* = c_3 \sum_1 \frac{c_1}{R_1^2} - \frac{1}{4} \sum_{i \neq 1 \neq n} \left(\frac{R_i}{R_1 R_n} \right)^2 \qquad (3\text{-}8)$$

So we notice that R^* is always negative in type-VIII spaces, whereas it can be positive or negative in type-IX depending on the relative difference among the various radii.

A remark should be made on using the system (3-7) in numerical integration. In highly anisotropic situations (vacuum stage) the energy density $\mathcal{E}$ in eq. (3-7b) is given by the difference between larger quantities, so it is preferable to use the energy balance equation:

$$\frac{d\mathcal{E}}{dt} = -3H(\mathcal{E} + p) - 9\eta_v H^2 + 4\eta_s (3H^2 - \mathcal{E} + \frac{1}{2} R^*) \qquad (3\text{-}9)$$

which indeed clearly displays the dynamics of the system once coupled with eq. (3-7a), leaving the relation (3-7b) as a constraint for the parameters specifying the initial conditions. In fact, because of the larger number of degrees of freedom allowed in the present case, the problem of initial conditions for numerical integration is now more complicated. Once $\mathcal{E}_{in}$ is determined by the solution class, several more parameters have to be chosen. In most of our calculation we adopted the following procedure: for a given world model we fixed $(R^*/H^2)_{in}$, $(R_1/R_3)_{in}$, $(R_2/R_3)_{in}$ and H_{in}, and set $H_{in} = H_{3\,in}$. Then eq. (3-8) determines the initial value of R_3, whereas the initial anisotropy was determined by the constraint (3-7b). Then, like in the flat space case, we allowed H_{in} to vary keeping the other parameter fixed: in this way one observes the influence of the initial anisotropy A_{in} on the properties of the world model. We remember that, because of the fluid scheme introduced in Chapter 2, the values of H_{in} are subject to the limits imposed by (2-6). The systematic change of the various parameters have been also investigated: in particular one observes the influence of the initial curvature repeating the whole procedure for several values of the initial curvature parameter $(R^*/H^2)_{in}$, ranging in magnitude from 10^{-50} to 10^{-16}. For some models we followed also the evolution in the subsequent plasma era making use of the adiabatic approximation.

The time evolution of a Bianchi IX model with $R^*/H^2 = 10^{-36}$ initially was reported in ref. [34]: Fig.(5)gives the curvature versus temperature dependence for several class A models. One notes that the

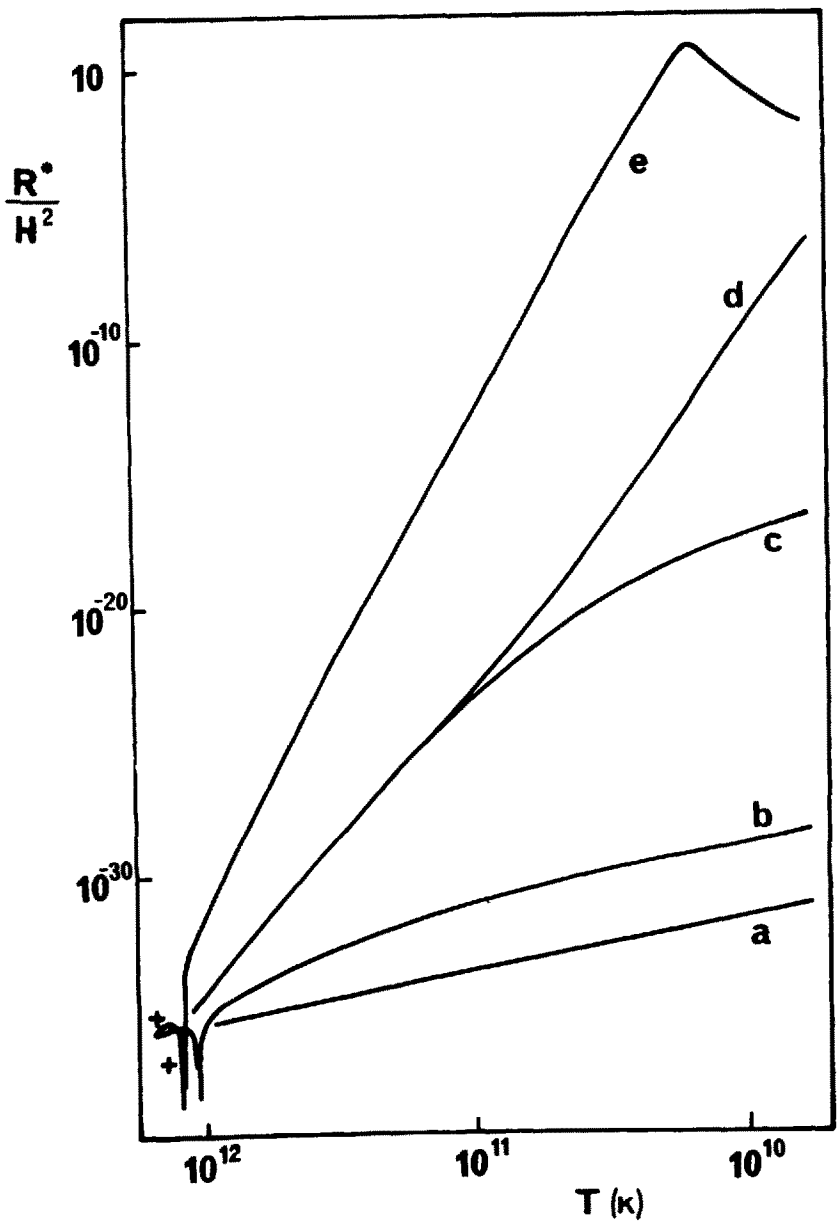

Fig. 5. The magnitude of R^*/H^2 versus the temperature for some class A models. The curves refer to the following values of H_{in} (cm^{-1}): (a) 3.6×10^{-4} (Friedman model), (b) 5×10^{-4}, (c) 10^{-5}, (d) 10^{-3}, (e) 15. Note that R^* is positive only in the branches labelled with the + sign. (Ref. [34]).

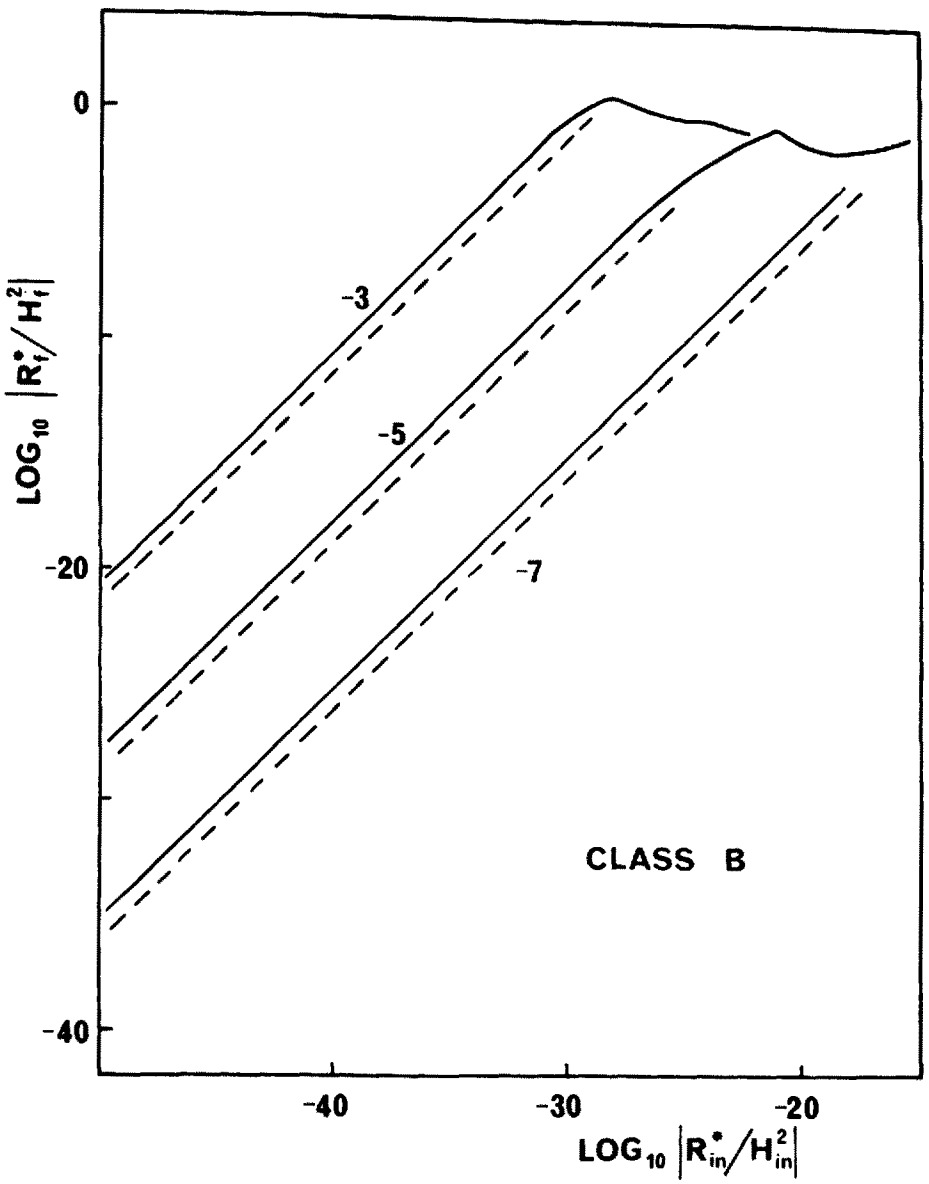

Fig. 6. The magnitude of the final curvature ratio $(R^*/H^2)_f$ versus $(R^*/H)_f$, for some values of H_{in}, in class B models. Note that R^*_{in} is positive for type IX (full lines) and negative for type VIII (dashed), whereas R^*_f is always negative. Each curve is labelled by the corresponding value of $\log_{10} H_{in}$. (Ref. [35]).

curvature quickly becomes negative since the scale factors become widely different due to the large expansion anisotropy. Afterwards the magnitude of R^*/H^2 increases very strongly due to viscous heating and becomes of the order of unity in the extreme cases. In these cases an oscillatory regime takes place near $T = 10^{10}$ K and goes on in the subsequent era (see curve (e)); however, as it will be shown below, the present day isotropy of the background imposes stringent limitation on the primeval curvature and the enormous increase of (R^*/H^2) reduces the set of plausible initial conditions.

The global properties of the world models may be discussed together in the type-VIII and type-IX cases. The most significant difference is the sign of curvature: type-VIII always displays a negative curvature, whereas type-IX can admit positive spatial curvature if the scale factors are rather close to each other. However, the cosmological evolution of anisotropic models typically leads to largely different scale factors; so, if $(R^*/H^2)_{in} > 0$, a sudden change in the curvature occurs in most models (however see ref. [35] in the case of axisymmetric models).

The evolution of the curvature is better understood looking at the quantity (R^*/H^2) which describes the bearing of curvature on the expansion rate. Generally speaking, anisotropic models are more influenced by curvature than strictly isotropic ones: the enormous production of radiating energy by viscous dissipation contributes to the enhancement of the curvature, which decreases much slower than in isotropic situations or even increases in highly anisotropic models. The magnitude of $|R^*/H^2|$ can increase up to 40 orders of magnitude (see Fig. (6)), so it can happen that $|R^*/H^2| \sim 1$ still at the lepton era. Therefore a single Kasner epoch is replaced by a curvature-dominated regime, switching the universe to another Kasner epoch, with the usual exchange in the sign of the Hubble parameters. For instance, in the lepton era of anisotropic models with $(R^*/H^2)_{in} = 10^{-16}$ we observed up to four different Kasner epochs. However, the entropy increase due to neutrino viscosity enhances the influence of radiation on the cosmological expansion. Thus, even if the oscillatory regime lasts a long time, the single stages are not strictly vacuum stages. In extreme cases this hybrid behaviour preludes a premature collapse of the universe.

In those models where the curvature remains small ($|R^*/H^2| < 10^{-2}$ throughout the lepton era) the cosmological evolution does not differ very much from the flat space case. A single Kasner epoch (no bounces) is replaced by a quasi-isotropic regime, as soon as the radiation con-

tent of the universe is no longer negligible. This situation occurs just at $T \sim 10^{10}$ K when the anisotropy of space is for all models $A \simeq 0.82$. Thus, owing to the viscous dissipation, a large set of initial conditions leads to the same state of the universe at neutrino decoupling and the residual anisotropy can be washed out in the subsequent expansion, whenever curvature permits.

The temporary optimism about isotropization is not however justified. In anisotropically curved spaces the curvature itself is strongly coupled to the expansion anisotropy so that, in general, high curvature prevents isotropization. As a matter of fact A is no more a decreasing function of time in models where $|R^*/H^2|$ is not negligible. This peculiar behaviour can be seen by inspection on Figs(7, 8) where models with $|R^*/H^2| \sim 1$ still at the lepton era are drawn. The anisotropy A reaches a minimum during the (high curvature) transition from one Kasner epoch to another, when all the Hubble parameters become temporarily positive, but new anisotropy is afterwards pumped on by curvature as a consequence of the $|R^*/H^2|$ decrease in the new Kasner epoch. In Fig.(7) the dashed lines give A_f as a function of H_{in} in type-IX spaces for two values of the initial curvature parameter, namely $R^*/H^2 = 10^{-32}$ (class A) and $R^*/H^2 = 10^{-36}$ (class B). The function clearly exhibits a characteristic behaviour for large H_{in}: it departs from the flat space value $A_f^{(o)} = 0.82$, reaching a minimum $A_f = 10^{-1} - 10^{-2}$; then it increases up to $A_f = 1.1$, and then drops slowly towards $A_f^{(o)}$. For comparison, R_f^* and A_f versus H_{in} are given in Fig.(8) for models where $R_{in}^* = 10^{-28} H_{in}$. One can see that, whenever $(R^*/H^2)_f$ is a decreasing function of H_i, the residual anisotropy is larger than $A_f^{(o)}$. This occurs for a set of models which exhibits two distinct Kansner epochs within the lepton era ($H_{in} \gtrsim 5 \cdot 10^{-3}$ cm^{-1}). For 10^{-4} cm^{-1} $\lesssim H_{in} \lesssim 10^{-3}$ the anisotropy is smaller than $A_f^{(o)}$; this refers to models where the universe is switching to a second Kasner epoch just at neutrino decoupling.

At any rate, if the residual anisotropies given by Figs.(7, 8) were adiabatically damped in the plasma era as in flat spaces, they would lead to an acceptable anisotropy of the background radiation today. But, as already remarked, the coupling of anisotropy and curvature does not allow a monotonic damping. An analytic description of this phenomenon may be given in the adiabatic regime. Then the differential Hubble parameters $\Delta H_k = H_k - H$ obey the equations:

$$\frac{d}{dt}\Delta H_k = -3H\Delta H_k + \frac{1}{3}\sum_{i\neq j\neq k}\left[\left(\frac{R_i}{R_jR_k}\right)^2 - \left(\frac{R_k}{R_iR_j}\right)^2\right] + \frac{c_3}{3}\sum_1 \frac{c_1}{R_1^2} - \frac{c_kc_3}{R_k^2} \qquad (3\text{-}9)$$

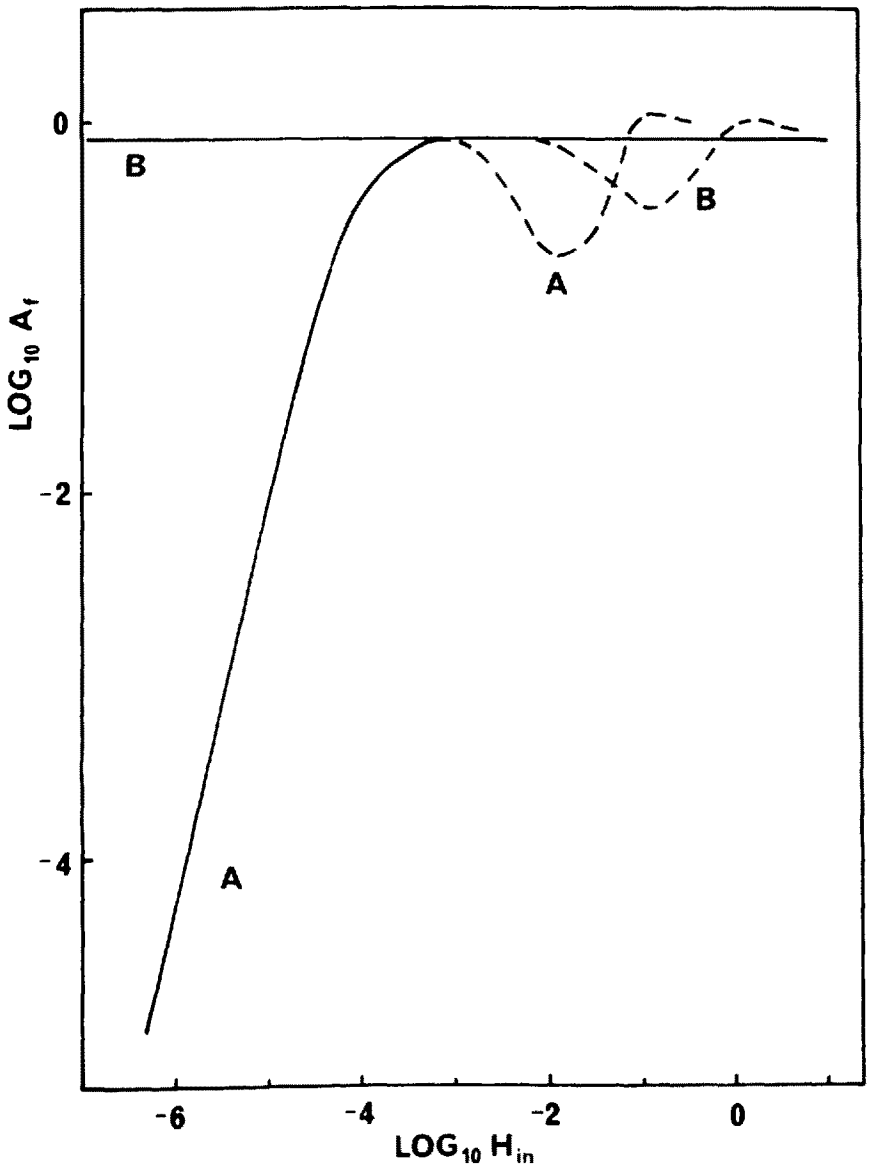

Fig. 7. The residual anisotropy A_f versus the initial Hubble parameter H_{in}. The full lines refer to quasi-flat models. The dashed line refers to type IX class A models with $(R^*/H^2)_{in} = 10^{-32}$ and class B with $(R^*/H^2)_{in} = 10^{-36}$. (Ref. [35]).

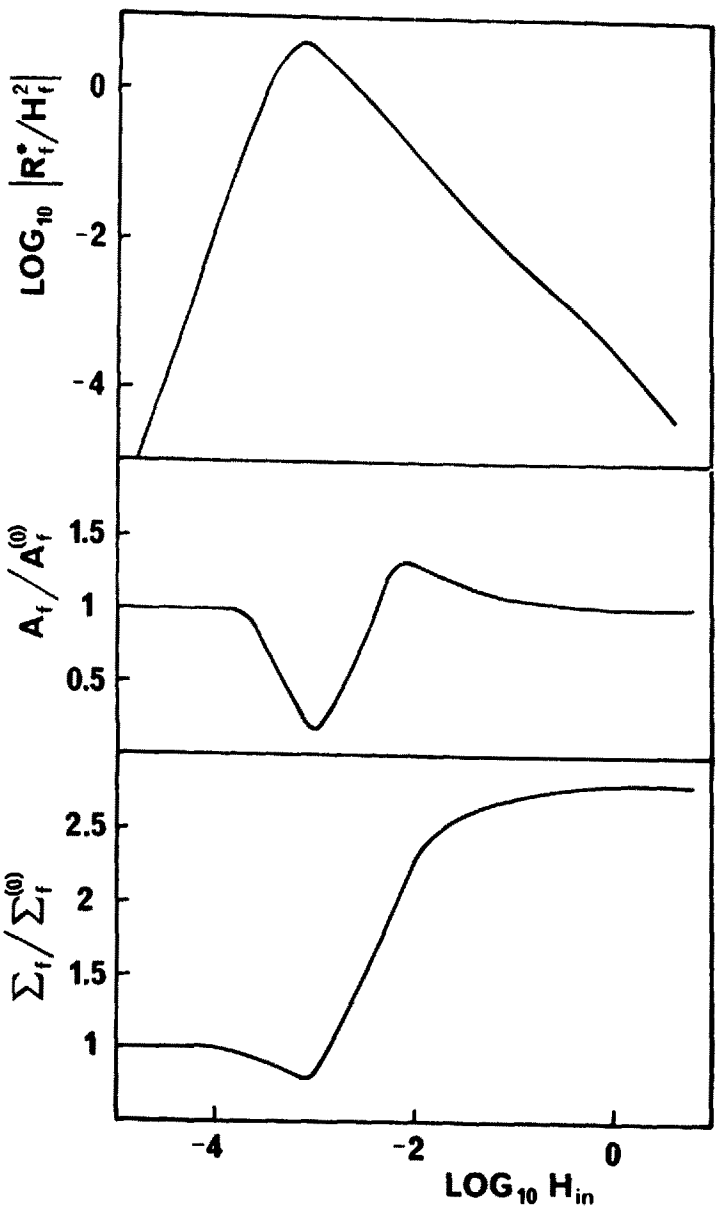

Fig. 8. Comparison of behaviours of R_f^*, A_f and Σ_f as functions of H_{in}. The curves refer to type IX, class B models with $(R^*/H^2)_{in} = 10^{-28}$ but their qualitative features are quite general. (Ref. [35]).

If $R_1 \gg R_2 \gg R_3$ the curvature R^* becomes

$$R^* = -\frac{1}{2}\left(\frac{R_1}{R_2 R_3}\right)^2 \tag{3-10}$$

so we obtain:

$$\frac{d}{dt}\Delta H_1 = -3H\Delta H_1 + \frac{4}{3}R^*$$
$$\frac{d}{dt}\Delta H_{2,3} = -3H\Delta H_{2,3} - \frac{2}{3}R^* \tag{3-11}$$

The solution of equations (3-11) is

$$\Delta H_1 = (R_1R_2R_3)^{-1}\left| a_1(t_o) + \frac{4}{3}\int_{t_o}^{t} R_1R_2R_3\, R^*\, dt \right|$$
$$\Delta H_{2,3} = (R_1R_2R_3)^{-1}\left| a_{2,3}(t_o) - \frac{2}{3}\int_{t_o}^{t} R_1R_2R_3\, R^*\, dt \right| \tag{3-12}$$

where $a_k(t_o)$ are constants. Thus the anisotropy can be given as an explicit function of time in the quasi isotropic stage ($A \lesssim 0.5$), where one knows $R_k \propto t^n$. As far as the integrals appearing in eqs. (3-12) are negligible, the anisotropy decays as in flat spaces

$$A = \text{const}\ R^{-6} H^{-2} \tag{3-13}$$

But the curvature term eventually prevails, then

$$A = \frac{8}{9}\left(\frac{n}{1+n}\right)^2 \left(\frac{R^*}{H^2}\right)^2 \tag{3-14}$$

where $n = \frac{1}{2}$ in the radiation dominated regime and $n = 2/3$ in the matter dominated regime. It is important to note that R^*/H^2 increases in magnitude like $t^2 R^{-2} \propto t^{2-2n}$ so that curvature eventually destroys the isotropy of the cosmological expansion. More precisely, if the matter dominated epoch started at a redshift $z = 10^4$, then $|R^*/H^2|$ has increased by 14 orders of magnitude since the end of the lepton era up to today. The present day limits on the quadrupole anisotropy of the cosmic background radiation imply that $A^{1/2} < 10^{-4}$ both at photon decoupling and at the present epoch. Consequently $|R^*/H^2|$ should be today less than 10^{-4} in closed and semi-closed models, unless strict isotropy is assumed since the beginning. Therefore we have the condition

$$\frac{R^*}{H^2} < 10^{-18} \tag{3-15}$$

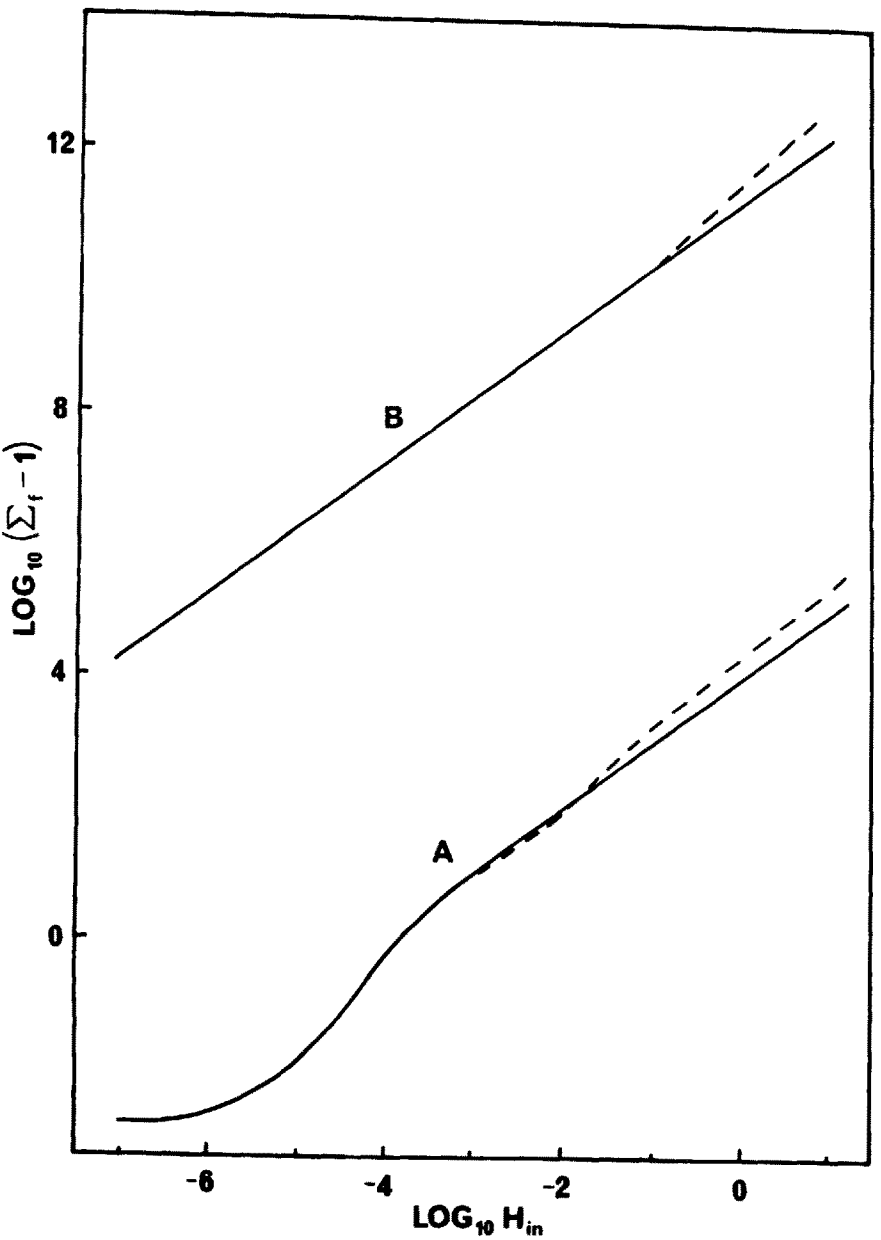

Fig. 9. The final entropy enhancement Σ_f versus the initial Hubble parameter H_{in} . The full lines refer to quasi-flat spaces. The dashed lines refer to class A models with $(R^*/H^2)_{in} = 10^{-32}$ and class B with $(R^*/H^2)_{in} = 10^{-36}$, as in Fig. 7. (Ref. [35]).

at the end of the lepton era. This is the most severe restriction in selecting the initial data which does lead to an acceptable universe today.

In models with $|R^*/H^2| < 10^{-2}$ throughout the lepton era, the viscous entropy production mechanism does not differ very much from the flat space, therefore eqs. (2-9) and (2-10) apply here also. The final entropy ratio $\Sigma_f^{(o)}$ (the superscript (o) means quasi-flat) as a function of H_i is reported by the full line in fig.(9). The class A curve exhibits a bend which is interpreted as the transition to quasi isotropic models. When $|R^*/H^2|$ is not small the behaviour of Σ_f is more complicated. With reference to fig. 8 we see that the entropy production is smaller than $\Sigma_f^{(o)}$ in models where $A_f < A_f^{(o)}$. We recall that a relatively small anisotropy is found in models just going to enter a new Kasner epoch at neutrino decoupling. The deficiency of dissipation confirms that the small anisotropy is merely due to a particular dynamical configuration, and is unrelated to the viscous damping. Conversely, when the universe enters the new Kasner epoch the anisotropy increases rather abruptly, and allows a higher dissipation. For models showing two fully developed Kasner epochs, the production of entropy is larger than $\Sigma_f^{(o)}$ by a factor $\simeq 2.7$. This is interpreted as the entropy enhancement due to a complete curvature bounce. In general, when the model displays N bounces, we found

$$\begin{aligned} \Sigma_f &\leqslant 2 \cdot 10^{12} \quad (2.7)^N \qquad \text{(class B)} \\ \Sigma_f &\leqslant 10^5 \qquad (2.7)^N \qquad \text{(class A)} \end{aligned} \tag{3-16}$$

However we have already remarked that, with high initial curvature only quasi-isotropic models are allowed; on the other hand only quasi-flat models are allowed if one wants the universe to be anisotropic at the beginning. Thus arbitrary dissipation is not permitted: looking at fig.(5) one sees that the fact of having considered models with initially $|R^*/H^2| = 10^{-36}$ rules out dissipation larger than 10^5 for class B and allows only $\Sigma_f \sim 1$ for class A solutions. Therefore the coupling between anisotropy and curvature can rule out many dissipative models, although curvature itself does not hamper dissipation.

Some conclusions can be drawn at this stage. According to our results one is forced to conclude that neutrino viscosity does not guarantee the isotropy of the universe in the closed and semi-closed case, even if viscous processes are very effective during the lepton era. The primeval anisotropy is substantially damped before $T \simeq 10^{10}$K, but new anisotropy is produced by curvature whenever $|R^*/H^2|$ is not

negligible. As a matter of fact, in many models with suitable curvature, eq. (3-14) applies well before the universe reaches its maximum expansion. In particular the universe may be now in a stage of increasing anisotropy, and therefore one might find a detectable anisotropy at the present epoch if the universe was exceedingly isotropic at the end of the lepton era. Realistic models must be nearly flat at $T \sim 10^{10}$ K, unless special conditions, i.e. strict isotropy, are assumed. Thus models with $|R^*/H^2| \sim 1$ during the lepton era are ruled out and the oscillatory behaviour described by Belinskii et al. [42] must concern more remote epoch. Moreover, we observe that realistic models must be almost flat at the present epoch, $|R^*/H^2| \leqslant 10^{-4}$, so that the matter density is close to the critical energy density $\mathcal{E}_c = 3H^2$ now. When the curvature is negative, the actual energy density, even in the closed Bianchi type IX, is less than $\mathcal{E}_c$. A positive curvature (then $\mathcal{E} > \mathcal{E}_c$) would be a priori unlikely at the present epoch, since it is allowed only for strictly isotropic models, for extreme axisymmetric models [35] and for accidental (i.e. temporary) configurations.

A strong production of entropy is not excluded, but the evolution of curvature imposes important limitations on the maximum Σ_f in realistic models with an assigned value of $|R^*/H^2|_{in}$. As found in flat space the production of entropy is $\Sigma_f \sim 10^{12}$ in the most dissipative case: the curvature of space can enhance this value if the universe is able to switch many times from one Kasner epoch to another still inside of the lepton era, but this possibility is not consistent with the present-day isotropy of the background radiation. Generally speaking, it seems to be easier for a general relativistic universe to produce entropy rather than to isotropize the expansion rate. In this connection two points are of the greatest importance:

a) The anisotropy-curvature coupling is an intrinsic property of the considered models and it is totally independent of the viscous mechanism. Although our approximations are unsatisfactory, the negative result concerning isotropization in curved models is unquestionable. Far from being a definite tendency of the cosmological evolution, the isotropic era of type-IX and VIII spaces is nothing else but a long span between two successive bounces.

b) The existence of earlier dissipative periods (in the very first quantum evolution) does not change the major conclusions. No matter how small the residual anisotropy can be, new anisotropy will be eventually produced if the universe is (anisotropically) curved. Only stringent limits on the initial curvature can eliminate this phe-

nomenon.

4. Bianchi type V Universes and no conclusions

The contradictory conclusions on the isotropization of flat and closed spaces compels us to check what neutrino viscosity can really offer to cosmologists by looking at the behaviour of open Universes. Here we briefly consider diagonal Bianchi type V spaces as a first example of (isotropically curved) open topologies. They constitute the natural generalization of the k = -1 Friedman models, and the Einstein equations for such spaces take the form:

$$\frac{dH}{dt} = -3H^2 + \frac{2}{R^2} + \frac{1}{2}(\varepsilon - p) + \frac{3}{2}\eta_v H$$
$$\frac{d\varepsilon}{dt} = -3H(\varepsilon + p) + 9\eta_v H^2 + 4\eta_s\left(3H^2 - \varepsilon - \frac{3}{R^2}\right) \qquad (4\text{-}1)$$

where, like in type-I spaces, the isotropic curvature allows a system of only two differential equations to be written down. Equations (4-1) can be integrated in the lepton era using the same numerical technique exposed in the preceeding chapters. The space curvature is now $R^* = -6/R^2$. The evolution of the curvature parameter R^*/H^2 has been found to be governed by the law:

$$\frac{R^*}{H^2} = \left(\frac{R}{H^2}\right)_{in}\left(\frac{H_{in}}{H}\right)^{4/3} \qquad (4\text{-}2)$$

Thus one observes that the enormous increase in the curvature found in closed spaces does not occur here. As a matter of fact, the law (4-2) describes the typical behaviour of vacuum stages and shows that the curvature is not influenced by the dissipative process. When high curvatures are reached ($R^*/H^2 \sim 1$) a saturation phenomenon takes place, so that R^*/H^2 approaches the limit value -6 and the universe enters the Milne epoch, where $\varepsilon \ll 3H^2$. Models in which this transition happens too early (inside of the lepton era itself for example, like in fig.(1), ref. [36]) should be however ruled out since they lead to a too low matter density at the present epoch. The damping of anisotropy is governed by the law $AR^6H^2 = \text{const}$, no matter how large the curvature is, even in the subsequent evolution. This implies that the coupling between anisotropy and curvature does not take place here. Moreover, since $R \propto t$ in the Milne stage, this law leads to a low residual anisotropy in high curvature cases. Generally speaking, the mechanism producing entropy seems not to be very sensible to the change

of curvature, and we obtain here the same results as in eqs. (2-9) and (2-10), except in cases where the Milne epoch begins inside of the lepton era. Then, owing to the fact that the expansion is faster, the production of entropy is reduced by a factor of order 3 with respect to quasi flat spaces. Consequently entropy enhancements as large as $\simeq 10^{12}$ are possible with suitable chosen initial anisotropies.

We can conclude that in type-V spaces neutrino viscosity could explain both the large radiation entropy in the universe and the degree of isotropy of the cosmic background radiation. They could surve as examples of models realizing the principles of chaotic cosmology. Of cource the crucial point for this conclusion is the absence of a curvature-anisotropy coupling.

It is then natural to ask oneself why the "Epicurean" program works so well in Bianchi type I and V and not in type IX and VIII. The reason seems to lie in the structure of the Riemann tensor, which is anisotropic in type IX and VIII and isotropic in type V (null in type I). Only the anisotropic part of the Riemann tensor is coupled to the expansion anisotropy. The conclusion, which is moreover supported by a preliminary investigation of the more general (i.e. anisotropic curvature) open models, can be stated as following: whenever the expansion anisotropy is purely kinematical, it is definitely damped by dissipative processes or at least by the adiabatic law $AR^6H^2 = \text{const}$. However, in the most general homogeneous world models (types VI, VII, VIII, IX) the anisotropy resides also in the intrinsic geometry of space, so it may be restored after any dissipative process, unless strong upper limits on the primeval curvature are posed. Selecting special initial conditions for an presently acceptable universe is then required and the philosophy of chaotic cosmology turns out to be thereby infirmed. It does apply only to quasi-flat spaces, whereas the presence of a (general anisotropic) curvature makes the isotropization of the cosmological expansion a difficult (if not impossible) task.

Going back to the introductory fig. 1, I believe that the isotropy of the universe is still to be explained, if theoretical cosmology could possibly ask the universe for such an explanation.

Acknowledgments

This work is based upon a research program which the author carried out in collaboration with R. Fabbri who kindly allowed myself to report about the common subject. I wish also to express my deep

gratitude to Martin Rees for warm hospitality and continuous support at the Institute of Astronomy.

References

1. C.W. Misner, 1967, Phys. Rev. Lett. 9, 533.
1968, Astrophys. J. 158, 431.
2. M.J. Rees, 1972, Phys. Rev. Lett. 28, 1669.
3. J.D. Barrow, 1977, Nature 267, 117.
4. J.D. Barrow, 1978, Nature 272, 211.
5. R.P.B. Partridge, this volume.
6. S. Weinberg, 1972, Gravitation and Cosmology, Wiley, New York.
7. T. Lucretius Carus, De rerum naturae.
8. E. Swedenborg, 1734, Principia Rerum Naturalis.
9. R. Penrose, to be published in Einstein Centenary Volume (Hawking and Israel eds. Cambridge Univ. Press).
10. J. Barrow and R.A. Matzner, 1977, Mon. Not. R. astr. Soc. 181,719.
11. B. Carter, in I.A.U. Symposium no. 63 (M. Longair ed.)
12. For a general review see: T. Criss, R. Matzner, M. Ryan and L. Shepley, in: 1975, General Relativity and Gravitation, eds G. Shaviv and N. Rosen, Wiley, New York.
13. A.G. Doroshkevich, V.N. Lukash and I.D. Novikov, 1973, Sov. Phys. JETP 37, 739.
14. M.A.H. MacCallum, this volume.
15. V.N. Lukash, I.D. Novikov, A.A. Starobinsky and Y.B. Zeldovich, 1976, Nuovo Cimento 35B, 293; Ya.B. Zeldovich, this volume.
16. L. Parker, 1977, in Proc. Symposium on Asymptotic properties of Space-Time, Plenum, New York.
17. A.G. Doroshkevich, Ya.B. Zeldovich and I.D. Novikov, 1967, JETP Lett. 5, 96.
1968, Sov. Phys. JETP 26, 408.
18. J.M. Stewart, 1969, Mon. Not. Roy.astr. Soc. 145, 347.
19. R.F. Carswell, 1969, Mon. Not. Roy. astr. Soc. 144, 279.
20. R.A. Matzner, 1969, Astrophys. Space Sci. 4, 459.
21. R.A. Matzner, 1969, Astrophys. J. 157, 1085.
22. S. Weinberg, 1971, Astrophys. J. 168, 175.
23. R.A. Matzner, 1971, Ann. Phys. 65, 438.
24. R.A. Matzner and C.W. Misner, 1972, Astrophys. J. 171, 415.
R.A. Matzner, 1972, Astrophys. J. 171, 433.
25. V.A. Belinsky and I.M. Khalatnikov, 1976, Sov. Phys. JETP 42, 205.
26. Z. Klimek, 1976, Nuovo Cimento 35B, 249.
27. S.L. Parnowskii, 1977, Sov. Phys. JETP 45, 809.
28. E.P.T. Liang, 1977, Phys. Rev. D16, 3369.
29. S.W. Hawking, 1966, Astrophys. J. 145, 544.
30. R. Hagedorn, 1973, in Cargese Lecture in Physics, E. Schatzman ed. Gordon and Breach.
31. N. Caderni and R. Fabbri, 1977, Phys. Lett. 69B, 508.
32. N. Caderni and R. Fabbri, 1977, Lett. Nuovo Cimento 20, 185.
1978, Nuovo Cimento 44B, 228.
33. N. Caderni, R. Fabbri, L.J. van den Horn and Th.J. Siskens, 1978, Phys. Lett. 66A, 251.
34. N. Caderni and R. Fabbri, 1978, Phys. Lett. 67A, 19.
35. N. Caderni and R. Fabbri, Preprint, Univ. of Cambridge.
36. N. Caderni and R. Fabbri, 1978, Phys. Lett. 68A, 144.
37. W.A. van Leeuwen, P.H. Polak and S.R. de Groot, 1973, Physica 63, 65.
38. W.A. van Leeuwen, A.J. Kox and S.R. de Groot, 1975, Physica 79A, 233;

A.J. Kox Ph.D. Thesis, Univ. of Amsterdam.
39. Th.J. Siskens and Ch.G. van Weert, 1977, Physica 86A, 80;
1977, Physica 87A, 369.
40. L.J. van den Horn and Th.J. Siskens, Physica (in press).
41. C.B. Collins and S.W. Hawking, 1973, Astrophys. J. 180, 317.
42. V.A. Belinskii, I.M. Khalatnikov and E.M. Lifshitz, 1970, Adv.
Phys. 19, 525.

COSMOLOGICAL MICROWAVE BACKGROUND BLACKBODY RADIATION AND FORMATION OF GALAXIES

Ya.B. Zel´dovich
Institute of Applied Mathematics, Academy of Sciences USSR, Moscow

It is now firmly established that at the early stages of the evolution universe was filled with hot dense matter. Discovery of the cosmological microwave background radiation by Penzias and Wilson (1965) and the observational indication that it has thermal spectrum was a great triumph of the big bang model of the universe (Gamow (1948), Peebles (1971)). Problems connected with observational investigation of the microwave background radiation were discussed by R.B. Partridge in his lectures. Here I would like to concentrate on physical processes which could leave some imprints on the spectrum and temperature of the microwave background radiation.

Let me first briefly review the basic properties of the electromagnetic radiation which is in equilibrium with matter.

1. Radiation in equilibrium with matter

In the homogeneous and isotropic world models (Friedman or Lemaître models) the microwave background radiation should have, in the first approximation, a blackbody spectrum (Peebles (1969))

$$n(\nu) = \begin{pmatrix}\text{occupation number}\\ \text{density of photons}\end{pmatrix} = \frac{1}{e^{\frac{h\nu}{kT}} - 1} = \frac{1}{e^{x} - 1}, \qquad (1)$$

where $x = \frac{h\nu}{kT}$. The total number density of photons is

$$N_\gamma = \frac{2}{(2\pi\hbar)^3}\int_0^\infty n d^3\vec{p} = T^3 \int n(x)x^2dx = 20\ T^3\ \text{cm}^{-3}, \qquad (2)$$

and the energy density is given by

$$\epsilon_\gamma = \frac{2}{(2\pi\hbar)^3}\int_0^\infty h\nu\, n d^3\vec{p} = 7.5\cdot 10^{-15}\ T^4\ \text{erg cm}^{-3} \qquad (3)$$

At $T = 3$ K, $N_\gamma = 540\ \text{cm}^{-3}$, $\epsilon_\gamma = 0{,}3\ \text{eV cm}^{-3}$, on the other hand the number density of electrons N_e and protons N_p is of the order of $N_e \sim N_p \sim 10^{-6}\ \text{cm}^{-3}$. It turns out that photons are the most numerous particles in the universe.

The Rayleigh-Jeans part of the spectrum is defined by the condition $x \ll 1$. In this region $n = 1/x$ and intensity of radiation is

$$F_\nu^{R-J}\left[\frac{\text{erg}}{\text{cm}^2\ \text{s Hz}}\right] = \frac{2kT}{\lambda^2}. \tag{4}$$

In the Wien region $x \gg 1$ we have $n = e^{-x}$ and

$$F_\nu^{W} = \frac{2h\nu^3}{c^2} e^{-\frac{h\nu}{kT}} \sim x^3 e^{-x}. \tag{5}$$

We may now ask, how and when during the expansion of the universe the equilibrium was established. In the approach to equilibrium emission and absorption of photons played an important role. These processes could be described by a kinetic equation involving Einstein's absorption coefficient A and emission coefficient B (Zel'dovich (1975))

$$\frac{\partial}{\partial t} n(\nu,t) = - An + B(1 + n), \tag{6}$$

here $B \cdot n$ and B represent correspondingly the contribution due to induced radiation and spontaneous radiation.

For a two level system A is proportional to the number density of unexcited states and B is proportional to the number density of excited states and as is well known $B = A\ e^{-x}$. Using this relation and taking $\frac{\partial n}{\partial t} = 0$ it is easy to obtain the equilibrium distribution

$$\frac{n}{1 + n} = \frac{B}{A} = e^{-x}, \tag{7}$$

from which it follows that

$$n = \frac{1}{e^x - 1}. \tag{8}$$

In plasma where both electrons and protons are present A and B are proportional to the product of number densities of electrons N_e and protons N_p,

$$B \sim \frac{N_e N_p}{x^3} e^{-x}, \qquad A \sim \frac{N_e N_p}{x^3}. \tag{9}$$

The kinetic equation (6) could be written in the form

$$\frac{\partial n}{\partial t} = - (A - B)n + B = - (A - B)(n - n_{eq}), \tag{10}$$

where n_{eq} is given by (8). The relaxation time $\tau = \frac{1}{A - B}$ is given by

$$\tau = \frac{1}{A - B} \sim \frac{x^3}{N_e N_p}. \tag{11}$$

In order to check if equilibrium existed at some epoch z_1 we introduce the notion of optical depth

$$\text{Optical depth} = D = \int_{z=z_1}^{z=0} \frac{dt}{\tau} \tag{12}$$

Let us remind that $z = 0$ corresponds to the present epoch. If a medium is transparent then $D < 1$.

During the radiation dominated epoch $z \sim t^{-1/2}$, so $dt = -\frac{1}{2}\frac{dz}{z^3}$ and $N_e N_p \sim (\text{density of matter})^2 \sim z^6$. Assuming $x = \text{const}$, we obtain

$$\int \frac{dt}{\tau} \sim \int \frac{dz}{z^3} \frac{N_e N_p}{x^3} \sim z^4. \tag{13}$$

It turns out that after annihilation of e^+, e^- pairs, at the beginning of radiation dominated epoch, when $z \sim 10^8$, D is of the order of unity for $x \sim 1$, i.e. for $h\nu \sim kT$, it means that the electromagnetic microwave background radiation was formed very early.

We should take into account, beside the absorption and emission of photons, another important process, namely the Compton scattering. In the nonrelativistic limit the cross section of the Compton scattering is

$$\sigma_T = \frac{8\pi}{3}\left(\frac{e^2}{m_e c^2}\right)^2 = 6.65 \cdot 10^{-25}\ \text{cm}^2. \tag{14}$$

The mean angle of scattering is close to $\pi/2$ and so if the initial distribution of photons was anisotropic, then after the first scattering the anisotropy will be reduced by 1/2, after the second scattering by 1/4, etc. At $z = 1000$ (roughly the recombination period) $N_e \sim 3 \cdot 10^3\ \text{cm}^{-3}$ and the characteristic time between two scatterings, which is also the characteristic time of smoothing out anisotropy, is

$$\tau \sim \frac{1}{\sigma_T N_e c} \sim 10^{10}\ \text{s}, \tag{15}$$

but the cosmological time corresponding to $z = 1000$ (characteristic time scale of expansion) is $t = 10^{13}$ sec. We conclude therefore that at $z = 1000$ any anisotropy is immediately smoothed out.

Let us now analyze how the spectrum changes due to scattering of photons with electrons. In the classical description (Thomson scattering) electrons oscillate with the same frequency as the incoming wave and the frequencies of incoming wave and scattered wave are equal. It is not so in the quantum case. An electron which was initially at rest introduces a shift in the wavelength

$$\lambda - \lambda_0 = \frac{h}{m_e c}(1 - \cos\theta), \tag{16}$$

and the relative frequency shift is therefore

$$\frac{\nu - \nu_o}{\nu} \sim -\frac{h\nu}{m_e c}. \tag{17}$$

(Astrophysicists use frequency $\nu = 1/P$ instead of $\omega = 2\pi\nu$, and therefore they use $h = 6{,}62 \cdot 10^{-27}$ erg sec.). Even in the classical case, if the electron has nonzero velocity (its change during scattering is neglected) then

$$\left|\frac{\Delta\nu}{\nu}\right| \leqslant \frac{v}{c}, \tag{18}$$

and the shift depends on the direction of velocity (Doppler effect). In the first order in v/c the Doppler effect results in broadening of the spectral lines (Fig. 1).

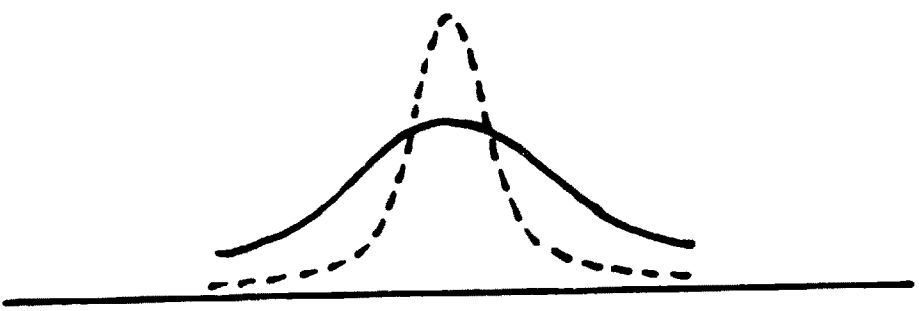

Fig. 1. First order Doppler effect broadens spectral lines

The shift of the whole spectrum is of the second order in v/c and

$$\left\langle \frac{\Delta\nu}{\nu} \right\rangle = \frac{v^2}{c^2} = \frac{kT}{m_e c^2}. \tag{19}$$

In thermal equilibrium the average Doppler red shift (19) is compensated by the quantum (Compton) blue shift (17). Obviously, the Planckian distribution of photons is not changed by the scattering if the electron temperature is equal to the temperature of radiation.

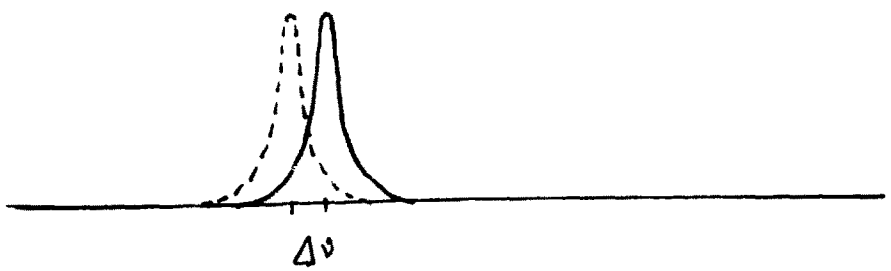

Fig. 2. Shift of the spectral line due to the second order Doppler effect

Let me now discuss the situation when radiation is not in equilibrium with matter.

2. Radiation not in equilibrium with matter

The non-equilibrium situation could be described by the Kompaneets-Weymann kinetic equation (Kompaneets (1957), Weymann (1965)). This equation can be used when

a) distribution of radiation is isotropic, and

b) $kT_e < m_e c^2$,

c) $h\nu \ll m_e c^2$ but the ratio of photons with $h\nu < kT$ and $h\nu > kT$ is arbitrary, when T_e is the temperature of electrons.

These conditions are of course satisfied for $z < 10^8$.

The Kompaneets-Weymann equation could be written in the following form

$$\frac{\partial n}{\partial t} = \frac{c\, h\, \sigma_T\, N_e}{m_e c^2}\, \frac{1}{\nu^2}\, \frac{\partial}{\partial \nu} \left\{ \nu^4 \left[\frac{kT_e}{h}\, \frac{\partial n}{\partial \nu} + n(n+1) \right] \right\} \qquad (20)$$

This equation has an important property, it does conserve the total number of photons, since

$$\frac{dN_\gamma}{dt} = \frac{d}{dt} \int n\, d^3 \vec{p} = \int \frac{\partial n}{\partial t}\, d^3 \vec{p} = \text{const} \int \frac{\partial n}{\partial t}\, \nu^2 d\nu\,, \qquad (21a)$$

but

$$\int \frac{\partial n}{\partial t}\, \nu^2\, d\nu = \frac{ch\, \sigma_T N_e}{m_e c^2} \int \frac{\partial}{\partial \nu} \left\{ \nu^4 \left[\frac{kT_e}{h}\, \frac{\partial n}{\partial \nu} + n(n+1) \right] \right\} d\nu = 0 \qquad (21b)$$

When $T_e \gg T$, the Kompaneets-Weymann equation reduces to

$$\frac{\partial n}{\partial t} = \frac{c\, \sigma_T N_e kT_e}{m_e c^2}\, \frac{1}{\nu^2}\, \frac{\partial}{\partial \nu} \left(\nu^4\, \frac{\partial n}{\partial \nu} \right) = cN_e \sigma_T \frac{kT_e}{m_e c^2}\, \frac{1}{x^2}\, \frac{\partial}{\partial x} \left(x^4\, \frac{\partial n(t,x)}{\partial x} \right) \qquad (22)$$

It is not so difficult to find a general solution of this equation. Here however we are interested only in limiting cases. For $x \ll 1$, $n_{R-J} = 1/x$, we obtain

$$\frac{\partial n}{\partial t} = cN_e \sigma_T \frac{kT_e}{m_e c^2}\, \frac{1}{x^2}\, \frac{\partial}{\partial x} (-x^2) = -2cN_e \sigma_T \frac{kT_e}{m_e c^2}\, n. \qquad (23)$$

Let us introduce new variable $y = \int cN_e \sigma_T \frac{kT_e}{m_e c^2}\, dt$; then we get

$$n_{R-J} = n_o\, e^{-2y}\,. \qquad (24)$$

Because $n_{R-J} = 1/x = kT/h\nu$ we have also that

$$T_{R-J} = T_o\, e^{-2y}\,. \qquad (25)$$

Using the general expansion for the energy density we obtain

$$\epsilon = \epsilon_o e^{4y}\,, \qquad (26)$$

where $\epsilon_o = a\, T_o^4$.

When $x \gg 1$, in the Wien region, from (22) we obtain

$$\frac{\partial n_W}{\partial t} = c\, N_e\, \sigma_T \frac{kT_e}{m_e c^2} x^2\, n_W \,, \tag{27}$$

and solving this equation we get

$$n_W = n_o\, e^{x^2 y} \,. \tag{28}$$

Comparing (24) and (28) and taking into account that $\frac{dN_\gamma}{dt} = 0$, we see that interaction of photons with hot electrons redistributes photons from the low frequency region of the spectrum to the higher frequencies.

Let us express our result in terms of measurable quantities. We usually measure T_{R-J} and ϵ but not T_o. From (25) and (26) we find that

$$\epsilon = \epsilon_o e^{4y} = a T_o^4 e^{4y} = a T_{R-J}^4 e^{12y} \,, \tag{29}$$

and when y is small

$$\epsilon = a T_{R-J}^4 (1 + 12y) \,. \tag{30}$$

The non equilibrium processes could substantially influence the spectrum of radiation and noticeably change the energy density of radiation.

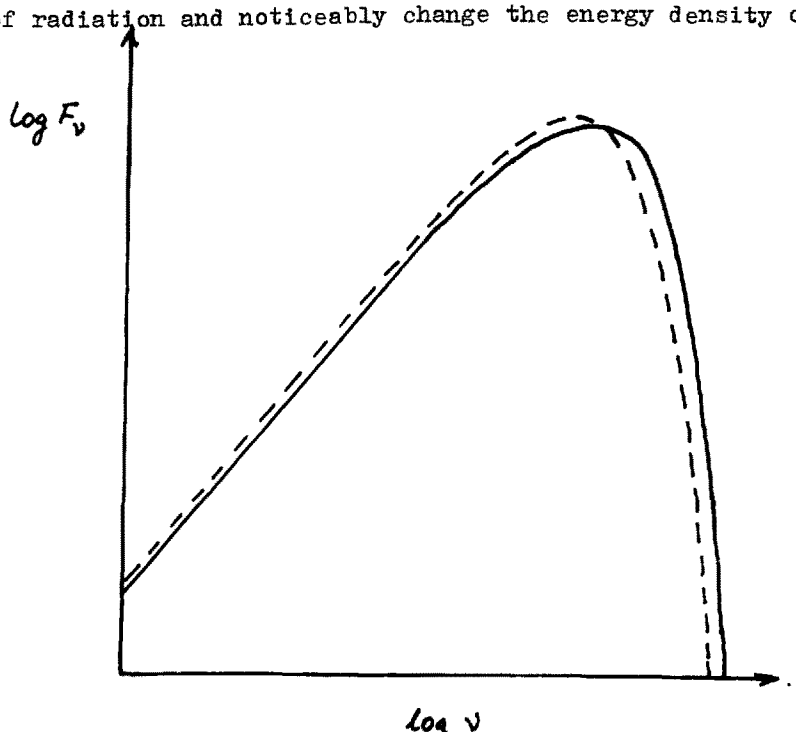

Fig. 3. Distortion of the initially blackbody spectrum (dashed line) by Compton scattering of photons on hot electrons

How does the temperature of electrons change due to interactions with radiation? In the general case, far from equilibrium, the Bremsstrahlung process is not efficient enough to establish equilibrium but collisions of electrons with themselves "maxwellize" the distribution. So we have that

$$\left(\begin{matrix}\text{relaxation time}\\ \text{of electrons}\end{matrix}\right) \ll \left(\begin{matrix}\text{relaxation time}\\ \text{of photons}\end{matrix}\right), \qquad (31)$$

which could be explained be "two reservoir" picture. Electrons approach the equilibrium much faster than photons, $\tau_e/\tau_\gamma \sim 10^{-8}$-$10^{-9}$.

When $N_e \ll 1$, the interaction of electrons with radiation is more complicated. This problem was investigated by Dreicer (1964), Peyraud (1968), and Zel'dovich and Levich (1970). Using the Kompaneets equation (20) and the definition of ϵ we get

$$\left(\begin{matrix}\text{Increase of energy density}\\ \text{of radiation in a given}\\ \text{volume}\end{matrix}\right) = \frac{d\epsilon}{dt} = \frac{4kT_e}{m_e c}\sigma_T N_e a h \int_0^\infty n\nu^3 d\nu -$$

$$- \frac{\sigma_T N_e}{m_e c} a h^2 \int_0^\infty n(n+1)\nu^4 d\nu = -\frac{d\epsilon_e}{dt} = \left(\begin{matrix}\text{Loss in energy density}\\ \text{of electrons in the}\\ \text{same volume}\end{matrix}\right) \qquad (32)$$

The stationary value of temperature of electrons can be obtained from the condition $\frac{d\epsilon}{dt} = 0$, which leads to

$$T_e^{stat} = \frac{h\int n(n+1)\nu^4 d\nu}{4k\int n\nu^3 d\nu} = \frac{c^2}{8k}\,\frac{\int F_\nu \left(F_\nu + \frac{2h\nu^3}{c^2}\right)\frac{d\nu}{\nu^2}}{\int F_\nu d\nu} \qquad (33)$$

This result could be also applied to plasma surrounding a quasar. Even a modest intensity of radiation at very small frequency can very efficiently heat the electron component of plasma, due to the appearance of ν^{-2} in the integral (33).

Let us now discuss a few examples of interest for cosmology. At the moment of recombination ($z = 1000$), $kT = E_I/-\ln(a_o^3 N_e)$ where a_o - the Bohr radius of hydrogen atom, $a_o = 5 \cdot 10^{-9}$ cm, $E_I = 13{,}5$ eV = = 160000 K, so T = 4000 K. If there was no reheating of gas after the recombination was completed, photons would not interact with bound electrons since that moment. Radiation and matter is no more in thermal equilibrium. But the cosmological expansion proceeds in such a way that it preserves the equilibrium spectrum even without any interaction between radiation and matter (Peebles (1969)).

This might be, however an idealized picture. If due to some violent processes after the recombination the plasma was reheated and ionized, this could have introduced a distortion in the spectrum of

the type discussed above (see Fig. 3). These distortions in principle could be observed.

In our previous considerations we introduced a parameter y, which describes distortion in the spectrum (comparing the bolometric energy density and the intensity in the long wave region). From the observational data it follows that $y \leqslant 0{,}02$, then $\epsilon \leqslant 1.2\, aT^4_{R-J}$ (see De Zotti, this volume). Let D be the optical depth of the medium. We have the relation $y = \int cN_e \sigma_T \frac{kT_e}{m_e c^2}\, dt = \frac{kT_e}{m_e c^2} \int cN_e \sigma_T dt = D \frac{kT_e}{m_e c^2}$.

On the other hand for D we have

$$D = 0.03 \left[(1+z)^{3/2} - 1\right] \Omega^{1/2} \tag{34}$$

where $\Omega = \rho/\rho_{crit}$ is the density parameter. If $\Omega = 1$, $z = 30$ then $D = 5$, and if $\Omega = 1$, $z = 10$ then $D = 1$. This result means that if the secondary ionization took place after $z = 10$, the universe would be transparent and spectrum of radiation would not be distorted.

Consider now the case of hot plasma and let the temperature of its electron component be 10^6 K. Then $kT_e/m_e c^2 \sim 1/5000$ and therefore due to Compton scattering the radiation will be very efficiently isotropized if z of the moment of ionization is greater than 10 (since then $D > 1$), but the scattering will not produce any noticeable distortion in spectrum as long as z of the ionization moment is less than 100 (since then $y \leqslant 0.006$, if $\Omega = 1$).

3. Processes in a rarefied plasma

When collisions between particles and Bremsstrahlung process (free-free emission) could be neglected but the Compton redistribution of photons takes place, then at the restricted equilibrium occupation number density is given by the Bose-Einstein (B-E) formula

$$n(\nu) = \frac{1}{e^{\frac{h\nu+\mu}{kT}} - 1} \tag{35}$$

where $\mu = \mu(N, \epsilon)$ is the chemical potential. This distribution describes an equilibrium state without absorption and emission. One can check this by inserting (35) into the Kompaneets equation and noticing that in the general case of $\mu \neq 0$, still $\frac{\partial n}{\partial t} = 0$. The distribution (35) has a shape different from the Planckian spectrum, which is the limiting case of $\mu \to 0$. Now when $h\nu/kT = x \ll 1$

$$n_{R-J} = \frac{1}{e^{\mu/kT} - 1} \tag{36}$$

(in the Planckian case we had $n_{R-J} = 1/x$), and

$$F_{\nu}^{B-E,R-J} \sim \nu^3 \tag{37}$$

(in the Planckian case we had $F_{\nu}^{P,R-J} \sim \nu^2$).

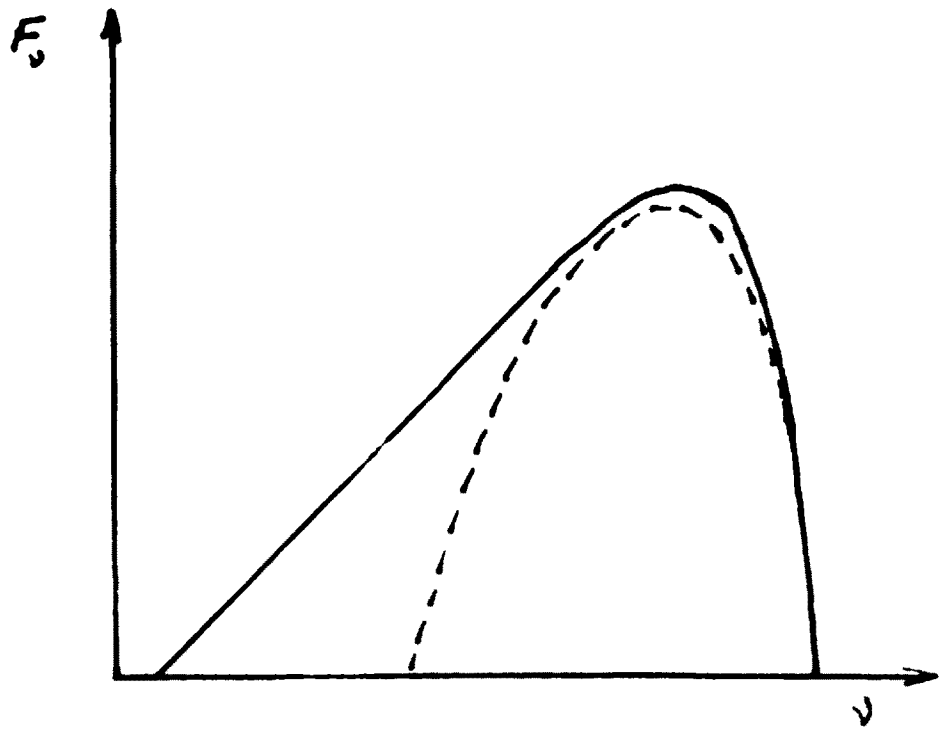

Fig. 4. Comparison of the Bose-Einstein distribution (dashed line) with the Planck distribution (solid line) of the same temperature

More detailed analysis, which takes into account the Bremsstrahlung process, leads to the conclusion that the distortion of the spectrum introduced by energy injection in the plasma at $z < 10^6$ should survive and it should not be smoothed out in the epoch when the Bose-Einstein formula with $\mu \neq 0$ describes distribution of photons. (Zel'dovich and Sunyaev (1969), Zel'dovich, Illarionov and Sunyaev (1972), Zel'dovich and Novikov (1975)).

The relaxation time depends strongly on frequency and there always exists some frequency band in the Rayleigh-Jeans part of the spectrum for which $\tau_{relaxation} \ll 1$, and the thermal equilibrium between photons and electrons is established. This exchange of energy pro-

duces a dip in the spectrum (see Fig. 5).

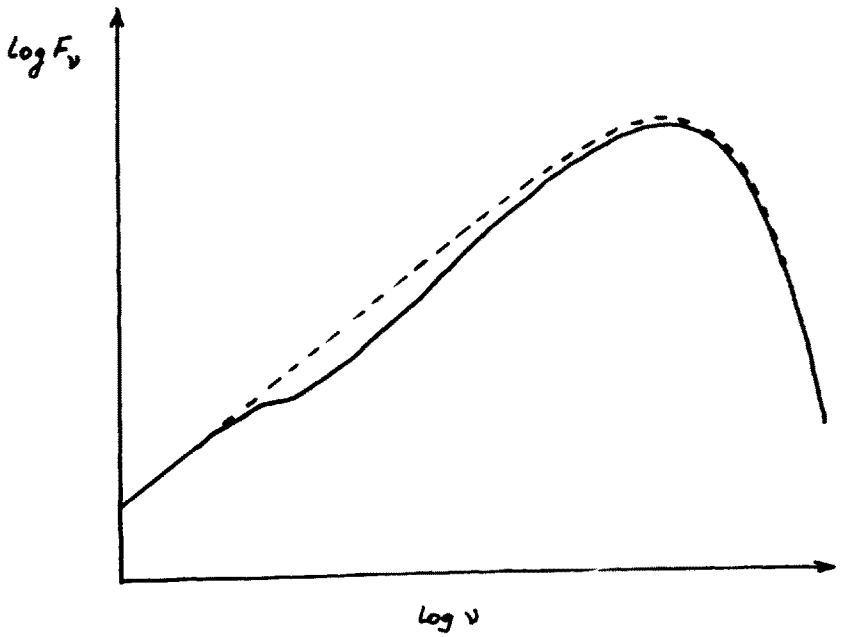

Fig. 5. Characteristic distortion of blackbody spectrum (dashed line) caused by early release of energy

The important conclusion is that in a real universe energy injected into plasma due to early damping of acoustic perturbations or evaporation of black holes or annihilation of antimatter, causes a definite distortion of the spectrum. Late energy injection and secondary ionization, due to quasar explosions etc. leads to quite different distortion of spectrum. Careful measurements of the spectrum are therefore very important for cosmology!

4. Small perturbations of the Friedman-Lemaître models

So far we were interested in the physical processes which could influence the cosmological microwave blackbody background radiation. In our considerations we assumed that the universe is homogeneous and isotropic and the perturbations were produced by energy injections

only. This is of course an idealization. We know that in small scale the universe is neither homogeneous nor isotropic, but in a sufficiently large scale (several hundreds of megaparsecs), according to the present observational data, the universe is surprisingly well described by the Friedman model (Longair (1974)). To be more realistic and to explain the existence of galaxies and their clustering we shall now discuss different modes of perturbations of the Friedman-Lemaître models.

We already noticed at the beginning that the number density of protons and other baryons is much smaller than the number density of photons. Since at the very early stages of evolution, when temperature was high enough, baryons and antibaryons were in thermal equilibrium, we have to assume that initially

$$B = (1 + 10^{-9})\ \overline{B} \tag{38}$$

where B and $\overline{B}$ are respectively the number density of baryons and antibaryons. Now in a fashionable theory of entropy perturbations, it is assumed that this excess is not uniform in space and at large scales

$$B = (1 + 10^{-9} \pm 10^{-11})\ \overline{B} \tag{39}$$

but at small scales the excess of baryons could be larger than 10^{-9}. Such situation would lead to entropy perturbations of the order of 1% (corresponding to the ratio $10^{-11}/10^{-9}$) after annihilation of baryon-antibaryon pairs. When the ratio $B/\overline{B}$ is exactly constant everywhere, other types of perturbations could be considered.

Let us look at the situation near the singularity, where the metric has an asymptotic form

$$ds^2 = dt^2 + a^2(t)\ g_{\mu\nu}(x)\ dx^\mu dx^\nu + b^2(t)\ k_{\mu\nu}(x) dx^\mu dx^\nu \tag{40}$$

Since near the singularity matter should be very hot, therefore for equation of state we can take the relation $p = 1/3\epsilon$, valid for an ultrarelativistic gas. From the Einstein's equation we can deduce that $a \sim t^{1/2}$, $b \sim t$ and

$$\varrho(t) = \frac{3}{32\pi G t^2} + \frac{\mathcal{P}}{t} \tag{41}$$

where ϱ is the density of matter and $\mathcal{P}$ is the scalar curvature calculated from the $g_{\mu\nu}$ part of the metric ($\mu,\nu = 1, 2, 3$) of the $t = \text{const}$ hypersurfaces. The first term in the expression for ϱ is the leading Friedmanian term and the second describes density pertur-

bations. The leading spatial part of the metric $g_{\mu\nu}$ could be decomposed into a flat, spherical or pseudospherical metric $\gamma_{\mu\nu}$ and a perturbation $h_{\mu\nu}$, so

$$g_{\mu\nu} = \gamma_{\mu\nu} + h_{\mu\nu} \tag{42}$$

If $h_{\mu\nu}$ is small, it corresponds to density perturbations and gravitational waves, which propagate on the isotropic and homogeneous background. If $h_{\mu\nu}$ is not small, non linear effects could play an important role. Sufficiently dense small regions could collapse and form small primordial black holes (PBH). The PBHs were first considered by Zel'dovich and Novikov in 1966. Recently formation of PBHs in an idealized case of spherical perturbations was studied numerically by Nadyozhin et al. (1977) (see also Carr and Hawking (1974) and Carr (1975)). As was shown by Hawking (1974), small black holes would evaporate producing bursts of X-rays. When $h_{\mu\nu}$ is not small in large regions then these inhomogeneities collapse and produce a massive PBHs. Observed isotropy of the microwave blackbody background radiation indicates that the initial perturbations were small on large scales.

In the theory of adiabatic perturbations one usually assumes that $h_{\mu\nu}$ is very small, $h_{\mu\nu} \sim 1\% - 0{,}1\%$. In a classical paper Lifshitz (1946) analyzed small perturbations of Friedman models. For the Fourier transform of $h_{\mu\nu}$ in the case of open models he obtained

$$h_{\mu\nu}(k) \sim \frac{\sqrt{3}\,\sin k\eta/\sqrt{3}}{k\eta}\; e^{ikx} . \tag{43}$$

where $d\eta = dt/a(t)$, $\eta(t) \sim t^{1/2}$ and k is the wave vector. $h_{\mu\nu}$ describes an oscillating perturbation of metric with decreasing amplitude. Density perturbations first increase and after some time oscillate. Formula (43) is valid for relativistic gas, and in particular for radiation dominated plasma i.e. when $z > 1000$. Density perturbations of masses smaller than the Jeans mass M_J ($M_J = \frac{4}{3}\pi v_s^3 \rho_0 (\frac{\pi}{G\rho_0})^{3/2}$) behave as acoustic waves. The Jeans mass just before the recombination is of the order of $10^{17}\ M_\odot$. Density perturbations of small masses are strongly damped, due to radiation induced viscosity. The smallest mass which survives through the radiation period until the recombination is called the Silk mass M_S and it is of the order of $10^{13}\ M_\odot$ (Silk (1968), Peebles and Yu (1970), Weinberg (1971)). If $M > M_J$ the perturbation is unstable and it will grow larger and larger.

During the damping process in the radiation epoch energy is injected from plasma to the radiation. As was already mentioned above this in principle could leave a noticeable imprint on the microwave

background radiation.

5. Adiabatic theory of formation of galaxies

Let us assume that the ratio of the number density of photons to baryons is determined by some basic physical theory (charge symmetry, baryon nonconservation etc.) and it is everywhere the same. We will consider a metric perturbations small enough to be treated in the linear approximation with a broad smooth featureless spectrum. After the radiation epoch, at the moment of recombination, the spectrum has a cutoff at $M_S \sim 10^{13\pm1}\ M_\odot$. After the recombination the Jeans mass drops drastically to $M_J \sim 10^5\ M_\odot$, so all the density perturbations which survived through the radiation epoch could now grow.

We will assume that initial conditions specified at the moment of recombination ($z = 1000$) are such that

$$\begin{pmatrix}\text{peculiar}\\ \text{velocity}\end{pmatrix} = \vec{u}(\vec{x}) \ll \begin{pmatrix}\text{Hubble velocity on}\\ \text{the same scale}\end{pmatrix}$$

$$(\text{density perturbations}) = \frac{\delta\rho}{\rho} \ll 1 \qquad (44)$$

$$\vec{\nabla} \times \vec{u} = 0 \qquad \begin{pmatrix}\text{no initial}\\ \text{vorticity}\end{pmatrix}$$

Now at $z = 0$, $\delta\rho/\rho \gg 1$. Therefore there existed a period $z_{rec} > z > z_1$ when the linear theory adequately described the evolution of perturbations, but this period ends before the present time. In that period for perturbations with masses much larger than $M_J \sim 10^5 M_\odot$ effects of pressure are negligible and therefore there is no dispersion $d\omega/dk = 0$. We can describe the density perturbation assuming that

$$\frac{\delta\rho}{\rho} \sim \sin kx\ e^{\omega t} \qquad (45)$$

then for ω we get

$$\omega = \pm (4\pi G\rho - \left(\frac{\partial p}{\partial \rho}\right)_s k^2)^{1/2} \approx (4\pi G\rho)^{1/2} \qquad (46)$$

Following Lifshitz (1946) and Bonnor (1957), who made similar calculations in the Newtonian theory, we can divide perturbations into two classes

transversal waves $\Rightarrow$ vortical perturbations
longitudinal waves $\Rightarrow$ density perturbations.

Let $\delta = \delta\varrho/\varrho$ and "g" stands for the growing mode and "d" for the decreasing mode, then in dust filled universe

$$\delta_g \sim t^{2/3} \qquad u_g \sim t^{1/3}$$
$$\delta_d \sim t^{-1} \qquad u_d \sim t^{-4/3} \tag{47}$$

for density perturbations and

$$\delta = 0 \qquad u_t \sim t^{-2/3} \tag{48}$$

for vortical perturbations. Therefore after some time the growing density modes will dominate and finally we get

$$\delta = t^{2/3}\, \delta_{init}(\vec{x})$$
$$\vec{u} = t^{1/3}\, \vec{u}_{init} \tag{49}$$
$$\vec{\nabla} \times \vec{u} = 0$$

for $z_{rec} > z > z_1$. But the linear theory is not adequate, since we have to consider situations when $\delta \gg 1$.

A natural way to formulate an approximate theory, exact in the linear region and good enough in the non linear regime, is to use the Lagrangian description (Zel'dovich (1970)). In that description the position of every particle (its Eulerian coordinate) $\vec{r}$ is given as a function of time t and the initial position (Lagrangian coordinate) of the particle $\vec{\xi}$.

The position vector $\vec{r}$ corresponding to the growing mode of perturbation is then given by

$$\vec{r} = a(t)\left[\vec{\xi} + t^{2/3}\,\vec{\psi}\,(\vec{\xi})\right] = t^{2/3}\left[\vec{\xi} + t^{2/3}\vec{\psi}\,(\vec{\xi})\right] \tag{50}$$

(Here we consider the case of $\Omega = 1$). The first term $a\cdot\vec{\xi}$ describes the Hubble expansion and the second $t^{4/3}\vec{\psi}$ describes the displacement of the particle from its unperturbed position. The perturbation caused by gravitational interaction is of the potential type, so

$$\vec{\psi} = \vec{\nabla}_{\vec{\xi}}\,\Phi \tag{51}$$

The velocity is given by

$$\frac{d\vec{r}}{dt} = \frac{2}{3}\,t^{-1/3}\vec{\xi} + \frac{4}{3}\,t^{-1/3}\,\vec{\psi} = H\,\vec{r} + \frac{4}{3}\,t^{1/3}\vec{\psi} \tag{52}$$

so that velocity perturbations are growing according to (49). Using (50) we can calculate the volume element

$$\begin{pmatrix}\text{Eulerian volume}\\ \text{element}\end{pmatrix} = d^3\vec{r} = t^2 \left\| \delta_{ik} + t^{2/3} \frac{\partial^2 \phi}{\partial \xi_i \partial \xi_k} \right\| = J d^3\vec{\xi} \qquad (53)$$

where J is the Jacobian. Therefore the density is given by

$$\rho(\vec{\xi},t) = \langle \rho(t_0) \rangle \; J^{-1} = \frac{\langle \rho\,(t) \rangle}{(1-\alpha t^{2/3})(1-\beta t^{2/3})(1-\gamma t^{2/3})} \qquad (54)$$

where we have used a coordinate frame in which J is diagonal and $\alpha, \beta, \gamma = -\frac{\partial^2 \phi}{\partial \xi_a^2}$ $a = 1,2,3$ are ordered so that $\alpha \geq \beta \geq \gamma$. Since ϕ can be determined from the initial conditions we can also calculate the value of α for every particle. The condition $1-t_i^{2/3}\alpha_i = 0$ determines a moment when the i-th particle will encounter infinite density. It is important that density becomes infinite because of vanishing of the denominator, i.e. due to contraction in one direction. Therefore the density becomes infinite on a surface on which the surface density is finite. At $t \longrightarrow t_i$, $\rho \longrightarrow \infty$ but the gravitational potential is finite and also the gravitational force acting on the particle is finite. Therefore at least qualitatively this approximate picture is correct. Let us call $\rho(\vec{\xi},t)$ as given by the formula (54) ρ_{TH}. Doroshkevich and Shandarin (1978a) made exact calculations of the density needed for the gravitational field to produce the motion of particles described by (50) and they obtained

$$\rho_{NEEDED} = \frac{\langle \rho \rangle \left[1 - (\alpha\beta + \alpha\gamma + \beta\gamma) t^{4/3} + 2\alpha\beta\gamma t^2\right]}{(1 - \alpha t^{2/3})(1 - \beta t^{2/3})(1 - \gamma t^{2/3})} \qquad (55)$$

We see that ρ_{TH} is a good approximation of ρ_{NEEDED}.

The main limitation of the theory is the fact that it gives a one dimensional singularity. If the initial cloud consists of stars they go through the formal surface of infinite density but when the initial cloud is just a cloud of hydrogen atoms then approaching the surface of infinite density atoms collide and form a shock wave. Therefore in the case of a cloud of gas we should get the following picture: in the general case, smooth initial perturbations will evolve, in the nonlinear stage, into ellipsoidal configurations with rapid contraction along the minor axis forming a pancake like structure. In the center density will grow and at some point shock waves will form and they will move out along the minor axis. The central part of the pancake will be therefore composed of a dense cold gas. The gas which passed through the shock wave will be hot and ionized. The pancake which corresponds to cluster of galaxies, will thereafter fragment into smaller pieces - galaxies, globular clusters and stars (Doroshke-

vich, Sunyaev, Zel'dovich (1974), Zel'dovich, Novikov (1975)).

The main prediction of adiabatic theory of formation of galaxies is that clusters of galaxies should form a net type structure. This prediction is tentatively supported by recent observations (Einasto, Joeveer (1978)). There are many N-body calculations which did not predict the net type structure. It is so because in those calculations the spectrum of perturbations is different from that which we have chosen in our analysis. More precisely, in the N-body calculations the fact that all perturbations with masses smaller than the Silk mass are damped out is not taken into account (the spectrum is not cutoff at short wavelengths). Numerical calculations of the evolution of perturbations in the two dimensional case done by Doroshkevich and Shandarin (1977) (1978b) also support the pancake picture and the predicted two point correlation function is in a good agreement with observational data. They are now analyzing the three point correlation function.

There is one important point which should be clarified, namely, from observations we know that galaxies rotate. On the other hand the Helmholtz-Kelvin theorem states that the potential forces could not produce angular momentum. This fact was the main psychological motivation of the vortex theory. In our theory however, the Helmholtz-Kelvin theorem is violated due to shock waves, and if the shock wave is not flat then even the initially potential, non vortical motion of gas will turn into vortical motion. So though initially $\vec{\omega} = 0$ after the gas passes through the shock wave $\vec{\omega} \neq 0$ (Doroshkevich (1972)).

We can also make some observational predictions. The gas in pancakes consisting mainly of hydrogen could emitt electromagnetic radiation at $\lambda = 21$ cm, this line could be redshifted to 2 - 3 m. It is now up to radioastronomers to find out if it really exists.

References

Bonnor, W.B., 1957, Mon. Not. R. astr. Soc. 117, 104.
Carr, B.J., 1975, Astrophys. J. 201, 1.
Carr, B.J., Hawking, S.W., 1974, Mon. Not. R. astr. Soc. 168, 399.
Doroshkevich, A.G., 1972, Astr. Zh. 49, 1221.
Doroshkevich, A.G., Shandarin, S.F., 1977, Astr. Zh. 54, 734.
1978a, Astr. Zh. 55, 1144
1978b, Mon. Not. R. astr. Soc. 182, 27.
Doroshkevich, A.G., Sunyaev, R.A., Zel'dovich, Ya.B., 1974, in Confrontation of cosmological theories with observational data, Ed. Longair, M., Reidel, Holland.
Dreicer, H., 1964, Phys. Fluids 7, 735.

Einasto, J., Joeveer, M., 1978, in The large scale structure of the universe, Ed. Einasto, J., Longair, M., Reidel, Holland.
Gamow, G., 1948, Nature 162, 680.
Hawking, S.W., 1974, Nature 248, 30.
Kompaneets, A.S., 1957, Sov. Phys. JETP. 4, 730.
Lifshitz, E.M., 1946, J. Phys. USSR 10, 116.
Longair, M., Ed., 1974, Confrontation of cosmological theories with observational data, Reidel, Holland.
Nadyozhin, D.K., Novikov, I.D., Polnarev, A.G., 1977, preprint Nr.347 Inst. Cosm. Research, Moscow.
Peebles, P.J.E., 1969, Amer. J. Phys. 37, 410.
1971, Physical cosmology, Princeton University Press, Princeton, N.J.
Peebles, P.J.E., Yu, J.T., 1970, Astrophys. J. 162, 815.
Penzias, A.A., Wilson, R.W., 1965, Astrophys. J. 142, 419.
Peyraud, J., 1968, J. Physique, 29, 88, 306, 872.
Silk, J., 1968, Astrophys. J. 151, 459.
Weinberg, S., 1971, Astrophys. J. 168, 175.
Weymann, R., 1965, Phys. Fluids 8, 2112.
Zel'dovich, Ya.B., 1970, Astron. and Astrophys. 5, 84.
1975, Sov. Phys. Usp. 18, 79.
Zel'dovich, Ya.B., Illarionov, A.F., Sunyaev, R.A., 1972, Sov. Phys. JETP 35, 643.
Zel'dovich, Ya.B., Levich, E.V., 1970, Sov. Phys. JETP Lett. 11, 35.
Zel'dovich, Ya.B., Novikov, I.D., 1966, Astr. Zh. 43, 758.
1975, Stroienie i evolucija vsielennoj, Nauka, Moscow.
1980, Structure and evolution of the universe, Chicago University Press, Chicago, Illinois.
Zel'dovich, Ya.B., Sunyaev, R.A., 1969, Astrophys. Space Sci. 4, 301.

COSMOLOGICAL ANISOTROPIES IN THE MICROWAVE BACKGROUND

R.B. Partridge
Haverford College, Haverford, USA

1. Introduction

The last twenty years were rich in discoveries which have revolutionized cosmology. Thanks to observations of quasars and radio sources, to nucleocosmochronology, X-ray astronomy, and the redetermination of the distance scale, cosmology is now an established branch of science, with a firm grounding in observations. Perhaps no discovery has had so decisive an impact as the cosmic microwave background (henceforward abbreviated cmb), first detected by Penzias and Wilson in 1965. Even before the cosmological origin of the microwave background was established (see Peebles, 1971; Weinberg, 1972; Partridge, 1975, for reviews), work was begun to make use of the observations to answer important cosmological questions. The achievements of theoreticians (some of whom are here at this school) are impressive in this respect; we know far more about the large-scale structure and evolution of the Universe than we did in 1965. These successes are even more remarkable when we consider that the observational data on which these results rest are really very few - we are talking about at most twenty observational papers.

In general terms, it seems to me that the cmb has proved useful for two reasons: the essential parameters (temperature and isotropy) can be measured to high precision, unlike many other cosmological quantities; and the cmb provides, in principle at least, a probe of the properties of the Universe at an epoch predating any structures we can now observe.

Let us consider two examples of the first point. The temperature of the radiation is known to be $T_0 = 2.8$ K, with an error of 10% or less. A simple projection of this result into the hotter denser past of the Universe permits a calculation of the rate of nucleosynthesis of light elements in the Big Bang (Wagoner, 1973). Good agreement is obtained with the observed abundance of He, thus giving strong support to the Big Bang Theory. In addition, these results provide an observational test for a vital cosmological parameter, the mean mass density of the Universe, based on the abundance of deuterium (Gott, Gunn, Schramm and Tinsley, 1974). It is important to note that an un-

certainty in the value for the temperature of the radiation as small as a factor of two would destroy the utility of the deuterium test. Likewise, accurate measurements of the large angular scale isotropy of the radiation can be used to limit sharply the number of possible anisotropic cosmological models - as already reported by Malcolm MacCallum. While the high degree of isotropy of the cmb does not prove that the Universe has a simple Friedman form, it does eliminate many other models which are consistent both with General Relativity and with other astronomical observations. (It should be mentioned that the abundance of light elements in the primordial material can fix even more restrictive limits for some Bianchi types - see Barrow, 1976 - but again, these results are sensitive to the value of T_0.)

The argument in this second example depends on the fact that the radiation we observe now was emitted in the distant past. This feature of the cmb comes more strongly to the fore when we consider the central question of my lectures -

What can observations of the cosmic microwave background tell us about the origin of structure in the Universe?

Potentially, careful observations of small angular scale anisotropies in the cmb can reveal a wealth of detail about the origin, nature and growth of the density perturbations which eventually grew into today's galaxies and clusters. I will develop the connection between density perturbations and observable temperature fluctuations in the cmb in somewhat more detail in a subsequent section. Before embarking on a detailed discussion, however, I want to emphasize that the promise of these observations has not yet been realized, since the observational material is not yet good enough. Just how galaxies and clusters formed remains one of the most important unanswered questions in cosmology.

2. Galaxy Formation

To make this point clearer, it is worth considering briefly and qualitatively the present view of galaxy formation and the problems this view encounters. In very rough form, one conventional theory holds that small density fluctuations early in the history of the Universe grow in amplitude until gravitational contraction, followed by fragmentation and star formation, takes place (for reviews, see Rees, 1971, Field, 1975, and Jones, 1976). Contraction commences at an epoch given by

$$t_f = \sqrt{\frac{3\pi}{32G\rho}}$$

where ρ is the density within the perturbation. Star formation presumably occurs in the interval $t_f - 2\ t_f$ (though more detailed models such as those of Larson, 1974, 1977, suggest later and more gradual onset of star formation). For reasonable values of ρ and typical cosmological models, values of t_f lie in the range 10^8–10^9 yrs., corresponding to a redshift of formation, $z_f \sim 3$–30. This model thus requires the existence of density perturbations of amplitude $\Delta\rho/\rho > 1$ and of galactic mass at an epoch like 10^8–10^9 yrs. As is well known, the classic and important work of Lifshitz (1946) has established that the amplitude of density perturbations grows very slowly in an expanding Universe in the linear regime:

$$\Delta\rho(z)/\rho_o \propto (1 + z)^{-1}$$

In an open, low density, Universe even this slow growth ceases at a value of the redshift given by $z + 1 \sim \Omega^{-1} \equiv \rho_c/\rho_o$, where ρ_c is the critical density given by $\frac{3\ H_o^2}{8\pi G}$ (Sunyaev, 1971). By themselves, these results would present no problems, since one could in principle start with arbitrarily small perturbations if one started at sufficiently large values of z. But in a hot expanding Universe there is a serious constraint because no growth in the amplitude of density fluctuations is possible until the cosmic background photons cease to interact with the matter content of the Universe, at $z \simeq 1000$. We will have a chance to look at this argument more carefully later; here I want only to emphasize that there is not much time for the density perturbations to reach an amplitude of unity or greater. If this conventional picture is correct, values of $\Delta\rho/\rho \simeq 3 \times 10^{-3}$–$3 \times 10^{-2}$ are required at $z \sim 1000$.

As we shall see it is not easy to account for density perturbations of order 10^{-2} at $z \simeq 1000$. Nor does it appear that we can ease our problem by imagining that galaxies form at much later epochs, so that the density perturbations have more time to grow. The few relevant observational results (Davis and Wilkinson, 1974; Partridge, 1974) suggest $z_f \geqslant 7$. Furthermore, late galaxy formation seems excluded in the presently-favored low density models by the argument given above (Sunyaev, 1971).

Of course it is always possible to abandon this picture. Indeed, there seems to be growing uneasiness about the conventional view which holds that all galaxy formation occurred at a single epoch in the past (see various papers in Tinsley and Larson, 1977). New models may be needed in this volume, Zel'dovich discusses in detail the theory of "pancakes" developed by the Moscow group. In this picture, the first

condensations to collapse gravitationally are far larger than galaxies - galaxy formation occurs later. Initial density fluctuations are still called for in this theory. In both this and the theory described previously, fragmentation is the primary mode of star and galaxy formation. The opposite approach is that of Press and Schechter (1974), who argue for gravitational condensation of preexisting smaller units, which generates condensations on larger and larger scales as time goes on. In this same vein is the suggestion of Rees (1977) that galaxies formed only after an initial generation of stars were formed. Since gravitational contraction (of collections of stars in this case) is still involved, perturbations of galactic mass are still required - but the epoch of galaxy formation can be later without violating the observational limits established by Davis and Wilkinson (1974).

As many authors have emphasized, we are very far from answers to the general questions of how and when galaxies formed. There is, therefore, considerable interest in the possibility that observations of the cmb will permit us to say something about the nature, scale and amplitude of protogalactic density perturbations.

3. The Last Scattering of cmb Photons

When we observe the cmb, we are seeing back to the surface of last scattering, the region (or epoch) at which the photons last scattered from matter. (One can think by analogy of observing the photosphere of the sun.) We can be certain that the surface of last scattering does not lie further away from us - or at an earlier epoch - than $z \sim 1000$, for at earlier times the matter content of the Universe was still ionized and the radiation interacted closely with the matter via the mechanism of Thomson scattering (Peebles, 1968). Thus the epoch at which the matter content of the Universe recombined to form neutral atoms is frequently taken as the surface of last scattering. Since the recombination of matter is not instantaneous, but rather occurs over a range of redshifts $\Delta z/z$ of order 0.1, the surface of last scattering is not sharply delimited. It might in fact be better to speak of a shell of last scattering.*

While $z \sim 1000$ is a firm upper limit on the redshift of the sur-

*Even though the recombination and consequent decoupling is not instantaneous, to first order the thermal spectrum of the cmb is maintained. This result is a consequence of the wavelength independence of the primary scattering mechanism, the Thomson process. See the contribution of Zel'dovich for details and modifications.

face of last scattering, there is a possibility that the photons last scatter at a much smaller redshift. If the process of primordial star formation or primordial galaxy formation is energetic, the matter content of the Universe may be reionized well after $z \sim 1000$, and Thomson scattering of the cmb photons may again become important. If reionization occurs, and if the optical depth is sufficiently high, the surface of last scattering will then be at a much lower redshift, and our view back to a redshift of 1000 will be blurred. How likely is this possibility? A short detour is required to find out. If there is substantial intergalactic "missing" matter in the Universe, that matter is presumably hydrogen gas. If so, it must be ionized to escape the very stringent observational upper limit on the density of neutral hydrogen (Gunn and Peterson, 1965). As Zel'dovich and Sunyaev (1969) have shown, the matter could not have remained ionized ever since the epoch corresponding to a redshift of 1000. Any intergalactic HII must have been ionized at a subsequent epoch. The required input of energy would presumably have been supplied by bright stars in primordial galaxies, by supernovae, by shock waves in collapsing "pancakes", or by the formation of protostars. If $\rho_{HII} \simeq \rho_c$, the energy requirements are substantial. Putting this difficulty aside, let us suppose that there exists in the Universe intergalactic ionized hydrogen of density equal to the critical density. Then the optical depth due to Thomson scattering reaches unity at a redshift of about 13 (Gunn and Peterson, 1965; modified to H_o = 50 km/sec per Mpc), and this would then become the surface of last scattering. If the density of ionized matter were substantially lower, as seems to be favored by the arguments of Gott et al (1974), then one could "see" somewhat further, to $z \simeq 16$. Now if this redshift is larger than the redshift at which primordial galaxies or stars form and release their energy, the Universe will remain transparent to $z \simeq 1000$ even if all the intergalactic material is ionized. In effect, the Universe becomes opaque again only if matter is reionized at a sufficiently early epoch. But recall that very early formation of bound systems presents a problem because of Lifshitz' argument. Finally, to complete this detour, let me remark that even if the surface of last scattering is at a lower value of the redshift than 1000, we may still be able to observe fluctuations in the microwave background, since it is reasonable to expect that the matter at the time of star or galaxy formation would be inhomogeneously distributed. In general, I will assume that there has been no reionization, so that the surface of last scattering is at $z \simeq 1000$.

4. Sketch of the Theory of Density Perturbations

With this outline in mind, let us now consider the theory or the origin, nature and growth of density perturbations more carefully.

First, it may be helpful to distinguish the various sorts of perturbations that we would expect in a hot expanding Universe. The first type is adiabatic perturbations in which both ρ and T are perturbed, so that the entropy, as measured by the ratio of photon number, is the same inside and outside the perturbed regions. For such perturbations $\Delta T/T = 1/3\ \Delta\rho/\rho$. These are discussed more fully by Zel'dovich here.

Another possibility is isothermal or entropy perturbations where the matter density is perturbed but the temperature is not.

A third class of perturbation is turbulent perturbations or vortices or "whirls" (Ozernoi and Chernin, 1967; Ozernoi and Chibisov, 1971; Ozernoi, 1974; also Anile et al, 1976). These are large-scale turbulent eddies in the coupled matter and radiation. After recombination and decoupling, the fossil eddies in the matter density undergo gravitational contraction as outlined above.

How would such perturbations arise in a hot Big Bang model? In my view, we are very far from any proper answer to this question. The problem is particularly acute in the case of adiabatic perturbations which are heavily damped before recombination. Thus initially large perturbations are required if $\Delta\rho/\rho$ is to reach recombination with the necessary amplitude. In our present state of ignorance, we may simply have to assert that the perturbations were there from the beginning, as initial conditions; to say metaphorically that it was God who separated the light from the dark.

The fate of density perturbations in the era before recombination depends both on their type and on their mass. The upper end of the mass spectrum we will have to consider is the mass contained within the light horizon of the expanding Universe at the epoch of recombination. Larger volumes were not causally connected. This upper limit is $M \sim 10^{19}\ M_\odot$.

First, let us consider adiabatic perturbations. A crucial variable is the Jeans mass (Jeans, 1928; see also Rees, 1971). Perturbations with masses grater than the Jeans mass can contract gravitationally, while smaller masses oscillate in amplitude. The Jeans mass itself varies throughout the epoch before and during recombination (Rees, 1971). Note that any density perturbation with mass $\lesssim 10^{18}\ M_\odot$ will have undergone a period of oscillation. A more precise value for

the maximum Jeans mass (Jones, 1976) is

$$(M_J)_{max} \sim 1.4 \times 10^{18} \frac{(\Omega h^2)^{-1/2}}{1 + 30\,(\Omega h^2)^{3/2}} M_\odot$$

Since the astrophysical systems of most direct interest - galaxies and clusters - have masses below this value, we need to ask what happens to perturbations which undergo oscillations after their mass has fallen below the Jeans mass. This question has been examined in considerable detail. Among others, Silk (1968, 1974) has shown that small scale adiabatic perturbations are heavily damped by photon diffusion and viscosity. Adiabatic perturbations with

$$M \lesssim 2 \times 10^{12}\,(\Omega h^2)^{-5/4}\, M_\odot$$

are damped out. The fact that the upper limit for radiative damping correspond roughly with galactic masses is intriguing (though one can argue as do Rees and Ostriker, 1977, that galaxies have the mass they do for different reasons).

Since in isothermal (entropy) perturbations the temperature is not perturbed, radiation damping of the kind described by Silk is not present. However radiation drag does prevent any growth in amplitude of isothermal perturbations whose mass is less than the Jeans mass. These will emerge at recombination with the same amplitude they had at the time the value of the Jeans mass equaled their mass. This is in sharp contrast with the case for adiabatic perturbations. It is worth noting that the Jeans mass is approximately equal to a galactic mass only a few years after the Big Bang. If one could detect fluctuations on the surface of last scattering, and if one could show that these are isothermal fluctuations (by considering the mass scale for instance), one would then have a sample of the density perturbation spectrum from a very early epoch.

The fate of turbulent (vortex) perturbations at and before decoupling is more complex. Briefly, turbulence on a particular mass scale may be damped, or even increased in amplitude as larger mass vortices decay. In addition, vortices may be heavily damped during recombination, when the vortex motions become hypersonic. These processes have been discussed in a series of papers by Ozernoi and his colleagues (e.g., Ozernoi, 1974), and reviewed by Jones (1976). While the exact results change from one paper to another, depending on the input physics, it appears that perturbations with masses $M \lesssim 3 \times 10^{12}\,(\Omega h^2)^{-7/2}$ are heavily damped. Assuming that the spectrum perturbations was initially Kolmogorov, one would then expect the maxi-

mum perturbation amplitude at approximately this value of mass.

Thus far, we have discussed perturbations in terms of the density contrast, $\Delta\rho/\rho$. The quantity of observational interest is the temperature contrast, $\Delta T/T$. As explained earlier by Zel'dovich, the relation between the two is simple only for massive adiabatic perturbations, where $\Delta T/T = 1/3\ \Delta\rho/\rho$. For masses $\lesssim 10^{15}\ M_\odot$, the observable ΔT is decreased because the individual perturbations are not optically thick at and just before recombinstion (Sunyaev and Zel'dovich, 1970). Thus any line of sight penetrates several perturbations and the temperature contrast is reduced. For these smaller perturbations as well as those with $M > 10^{15}\ M_\odot$, fluctuations are produced primarily by Doppler shifts caused by mass motions within the perturbations:

$$\Delta T/T = \int \frac{v(z)}{c} \cos\theta e^{-\tau(z)} d\tau(z)$$

where θ is the angle between $\vec{v}$ and the line of sight, and $\tau(z)$ is the optical depth for Thomson scattering (see Sunyaev, 1977a). For masses $\gtrsim 10^{12}$-$10^{13}\ M_\odot$, the expected values of the temperature fluctuation $\Delta T/T$ are $\sim 10^{-4}$ (Doroshekevich et al, 1977a). To be more precise, for a value of $\Delta\rho/\rho$ at $z = 1000$ of 10^{-2}, we expect a maximum value of $\Delta T/T$ of $\sim 2 \times 10^{-4}$ on an angular scale of $\sim 5'$.

One expects temperature fluctuations of approximately the same magnitude to be produced by the same effect by isothermal perturbations (Zentsova and Chernin, 1977). The only difference to note is that perturbations on mass scales below 10^{12}-$10^{13}\ M_\odot$ may be present, since radiation damping does not affect small isothermal perturbations as it does small adiabatic perturbations. Hence if perturbations are detected with masses of $\sim 10^{12}\ M_\odot$ or below, it would suggest they are isothermal in nature.

5. Observational Parameters

Let us now assess and summarize these results from an observational point of view, as Sunyaev (1978) did in his recent reviews in Tallinn. Several mass scales may be of particular interest. I have also added approximate values for the corresponding angular scale, making the assumption that the surface of last scattering is at $z = 1000$. Of course these values of θ will also depend on the cosmological model assumed. In those cases where estimates of $\Delta T/T$ have been made, they are also given.

$M \gtrsim 10^{19}\ M_\odot$; $\theta \gtrsim 2^\circ$. Such large perturbations were not inside

the light horizon before recombination; that is, these regions were not causally connected until after the epoch of $z \sim 1000$. Hence any perturbation observed on such a large angular scale must have been present ab initio.

$10^{19}\ M_\odot \geqslant M \geqslant 10^{15}\ M_\odot$; $2^\circ \gtrsim \theta \gtrsim 6'$. It appears that there are no strong concentrations of mass in this range (see the work of Peebles and his group, discussed in this volume by Dautcourt). It therefore seems reasonable that $\Delta\rho/\rho$ for such masses was not greater than 10^{-3} at recombination. On the other hand, it is also reasonable to expect some fluctuation on this scale if fluctuations on smaller scales were present - presumably the perturbation spectrum did not have a sharp cutoff at $10^{15}\ M_\odot$. Indeed, the simplest assumption one can make about the initial mass spectrum of the density perturbations is a power law, $\Delta\rho/\rho \propto M^{-2/3}$, in which case $\Delta\rho/\rho \sim 3\,\Delta T/T \sim 10^{-5}$–$10^{-4}$ might be expected (see, for instance, Sunyaev, 1978). Even substantially larger values of $\Delta\rho/\rho$ would be consistent with the observed clustering of matter on large scales, as discussed by Peebles and his colleagues.

$10^{15}\ M_\odot \geqslant M \geqslant 10^{12}\ M_\odot$; $6' \gtrsim \theta \gtrsim 1/2'$. This mass range includes the mass of clusters of galaxies. The Universe today is quite clearly inhomogeneous on the mass scales in this range. Therefore, as Boynton (1978) has emphasized, density perturbations must have been present in the past (figure 1). For any given value of $\Omega = \rho_0/\rho_c$, we can even estimate the amplitude of the perturbations required at recombination ($z = 1000$). Thus this range of masses (or of angular scale) offers the best chance of forcing a confrontation between theory and observation.

Detailed predictions of the amplitude of temperature fluctuations are provided by Sunyaev (1978). These calculations suggest maximum values of $\Delta T/T \sim 2\text{–}6 \times 10^{-5}$ at 5′–10′. Silk and Wilson (1979) have recently used the properties of rich clusters and the work of Seldner and Peebles (1977) to predict $\Delta T/T \sim 1.5 \times 10^{-3}$ for proto-cluster fluctuations. This value does not appear to take into account the reduction in ΔT which we expect because the fluctuations have optical depth < 1. Realistically, Silk and Wilson's work suggests $\Delta T/T \sim 1\text{–}2 \times 10^{-4}$, on angular scales of 1/2′–20′.

$M \leqslant 10^{12}\ M_\odot$; $\theta \leqslant 30''$. If fluctuations on these small angular scales are found, they must presumably be either turbulent or isothermal in nature. Towards the upper end of this range is the mass of a typical galaxy $\sim 10^{11}\ M_\odot$. We again expect $\Delta T/T \lesssim 10^{-4}$.

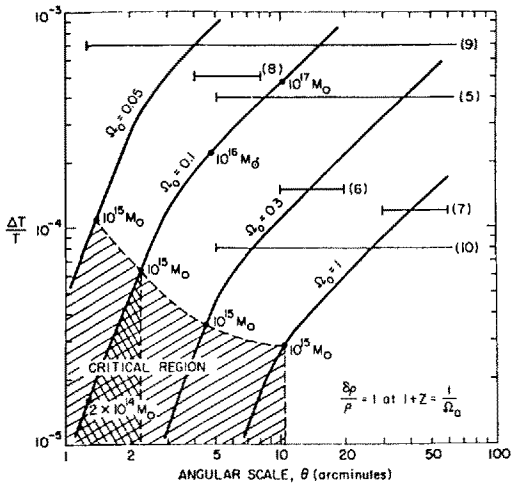

Fig. 1. Small angular scale fluctuations calculated from the work of Sunyaev and Zel'dovich (1970), for various values of Ω. The horizontal bars represent published or reported upper limits on $\Delta T/T$ on various angular scales. See Table 1 for corrected values, and for identifications. (From Boynton, 1978.)

The final quantity of interest to observers may be roughly called the geometry or morphology of the fluctuations. Let me make this point by asking a series of questions - to most of which answers are lacking. Can we expect the temperature fluctuations to have approximately radial symmetry? The work of Zel'dovich and his group suggests rather a chaotic structure, without neat, isolated, temperature fluctuations. In the case of vortex fluctuations, I would expect regions of positive ΔT to be immediately adjacent to regions of negative ΔT. Next, are perturbations of a particular angular scale randomly distributed on the sky, or ought we to expect higher order clustering? Answers to these questions - and indeed any detailed predictions of the geometry and amplitude of fluctuations - would be of considerable assistance to observers.

6. Other Possible Sources of Temperature Fluctuations

Thus far, I have restricted my discussion of temperature fluctuations to those produced by perturbations present at $z \sim 1000$. There are many other sources of temperature fluctuations in the cmb which I should mention, even if I cannot treat them in detail. In some cases, these additional mechanisms might produce values of ΔT larger than

the cosmological fluctuations. More detailed calculations of these effects would therefore be useful. There are, it seems to me, some interesting problems in this area, waiting for solution.

First, as Dautcourt (1969) has pointed out, long wavelength gravity waves can produce ΔT fluctuations. See also Doroshkevich at al (1977b).

Next are fluctuations which might be described as cosmogonic - fluctuations arising from the formation of gravitationally bound systems at $z \ll 1000$. A burst of star or galaxy formation which reionizes the intergalactic medium (see above) would produce temperature fluctuations if the heating were irregular (Sunyaev, 1977b). Shock waves in "pancakes" might produce ΔT fluctuations with the filimentary* or honeycomb structure shown in Zel'dovich (1978). This filimentary structure, if detected, would permit us to distinguish cosmogonic fluctuations from primordial ones. There is also the possibility, raised by Fabian and Rees (1979) of scattering at even lower redshifts ($z \sim 1$) in gaseous protoclusters. The primary mechanism suggested is inverse Compton scattering (see Sunyaev and Zel'dovich, 1972), which would produce "cool spots" in the cmb with $\Delta T \sim 10^{-3}$ K on angular scales of $\sim 1'$.

Finally, I should mention the obvious point that the presence of weak, unresolved, but nearby radio sources will also produce fluctuations in the observed intensity of different regions of the sky, quite unconnected with cosmological processes. This problem is discussed further below.

7. Basic Radio Astronomy

In the preceding pages, I've tried to show how observations of small-scale anisotropies could lead to the solution of important cosmological problems. We have seen that sensitivities of $\Delta T/T < 10^{-3}$ will be required if the observations are to be of use. Even greater sensitivity would be desirable. Can we do it?

Over a limited range of angular scales and wavelengths, the answer is yes. Indeed, there have already been published several investigations reaching well below the limit $\Delta T/T = 10^{-3}$. With one important exception, this research has employed conventional radio as-

*Since we are talking about second generation processes, one could also say "filamentary" - a comment no one who was not at Jodłowy Dwór will understand.

stronomical techniques, so it is appropriate to begin our consideration of the observations by reviewing some basics of radio astronomy.

To start at the very beginning, a radio telescope consists of a single receiver mounted at the focus of some sort of antenna, as shown in figure 2.

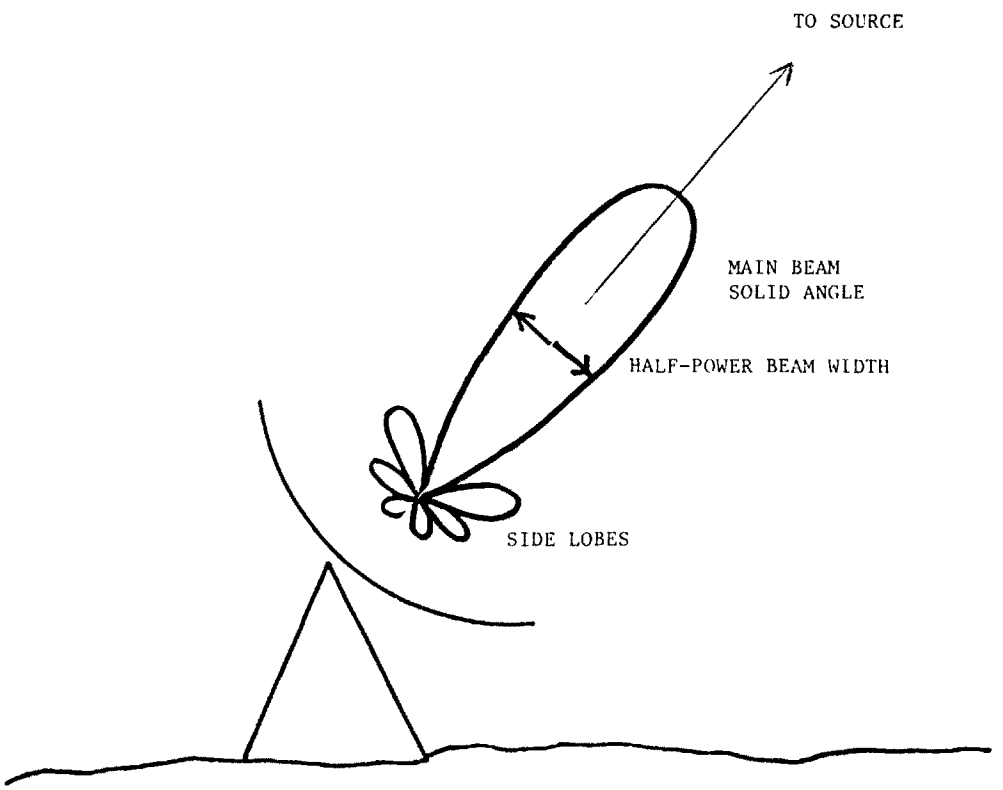

Fig. 2. The response of the telescope as a function of angle measured from the axis of symmetry is shown. The half-power beam width of the main beam solid angle is indicated.

Because of diffraction, these telescopes are sensitive not just to electromagnetic radiation incident exactly on axis, but also to radiation coming from a small solid angle around the axis. While most of the response of the telescope is concentrated in a small solid angle around the central axis (called the "main beam solid angle"), higher order diffraction maxima (referred to as "side lobes") can diffract

radiation from the ground into the receiver. We will see that side--lobe radiation can present some problems when making measurements at the high sensitivities which we require. One way to reduce the problem of side-lobe radiation, and also to reduce errors caused by random changes in the sensitivity of the receiver, is to make a differential measurement. One switches the receiver, electrically or mechanically, between two closely adjacent regions of the sky, one of which contains the source of interest, the other of which is nominally empty or "blank". Since one is in effect looking alternately at two separate small solid angles in the sky, this technique is referred to as beam switching. Finally, while one cannot eliminate side-lobe radiation from the ground entirely, one can at least limit changes in its intensity by making observations with the telescope fixed, so that the diffraction pattern does not change in time. Scans made of the sky as the celestial sphere rotates past a fixed telescope are referred to as drift scans. Many of the observations I will discuss later employ both beam switching and drift scan techniques.

Now let us consider the sensitivity of radio astronomical measurements. Ignoring for the moment external sources of noise, the limiting sensitivity of a radio telescope is fixed by the Johnson noise generated in the receiver (in the mixer of a superheterodyne receiver or in the various stages of amplification). This noise is traditionally described in terms of a temperature, the receiver noise temperature, such that $kT_R = W$, where W is the noise power per unit bandwidth. The actual noise temperature of a real radio telescope observing the sky through the Earth's atmosphere will be somewhat larger, for a variety of reasons which I will skip over. The actual system noise temperature T_s will exceed T_r by tens or hundreds of degrees depending on the system and the wavelength.

Given a system with noise temperature T_s, what is ΔT_m, the minimum r.m.s. temperature fluctuation in the sky which can be detected? It is

$$\Delta T_m = \frac{T_s}{\varepsilon_a \varepsilon_b \varepsilon_s \sqrt{\Delta\nu t}} \qquad (1)$$

where $\Delta\nu$ is the postdetection bandwidth of the receiver (generally fixed by the first stage of the amplifier), and t is the total integrating time. In this equation, I have used ε_a to represent the aperture efficiency (including the reflectivity of the surface of the antenna), and ε_b as the beam efficiency, which we may define as the following ratio, using S to represent the flux density or signal recorded:

$$\varepsilon_b = \frac{S\ (\text{point source on axis})}{S\ (\text{similar source filling total solid angle})}$$

Diffraction losses and any structural irregularities in an antenna will cause ε_b to be less than unity. Together $(\varepsilon_a \varepsilon_b)$ measures the efficiency of response to a source small in comparison to the main solid angle. Typical values for $(\varepsilon_a \varepsilon_b)$ are approximately 0.5. Finally, ε_s is a parameter which takes into account the mode of beam switching employed. If ordinary beam switching is employed, with square wave modulation, $\varepsilon_s = 1/2$.*

Corresponding to ΔT_m is the minimum detectable flux density (units of 10^{-26} watts/m^2 Hz), ΔS_m, given by

$$\Delta S_m = \frac{2k\Delta T_m}{A}$$

for ΔT_m as defined above, where A is the geometrical area of the telescope.

To be concrete, let me work out a specific example. The 11-meter millimeter wave telescope at the National Radio Astronomy Observatory in Tucson has a receiver with $T_r \sim 520$ K and $\Delta\nu = 10^9$ Hz. The system temperature is somewhat higher than T_r, typically $T_s = 570$ K. For this antenna at $\lambda = 9$ mm, $(\varepsilon_a \varepsilon_b) = 0.4$, and for beam switching $\varepsilon_s = 0.5$. For a 10 second integration, we find $\Delta T_m \sim .03$ K for a single point on the sky. To reach $\Delta T = 10^{-3}$ K = 1 mK would require ~ 3 hours of observation. To search for fluctuations, of course, one needs to observe a statistically significant number of independent points. Fairly obviously, one is talking about days of work at a sensitive radio telescope. The hope for the future lies either in lowering T_s, or, more likely, markedly increasing $\Delta\nu$. Nevertheless, with patience, instruments available today should be able to reach $\Delta T/T \lesssim 10^{-4}$, provided that other sources of random and systematic error can be kept low. Let us now turn to the question of errors, and the limits they set on our searches for fluctuations in the cmb.

For all of the searches for small scale anisotropy in the cmb which have thus far been published, the limiting factor appears to have been the system noise discussed above. This will probably continue to be true at short wavelengths. Receiver noise increases sharply as frequency rises; at present the technology of high frequency receivers, and the deterioration of ε_a at high frequencies, fixes a practical lower limit on the wavelength at which ordinary radio ob-

*For further details, consult a text on radio astronomy, such as Kraus (1966) or Shklovskii (1960).

servations can usefully be performed: $\lambda \geq 3$ mm.

In addition, high frequency measurements are more susceptible to noise introduced by reemision from the earth's atmosphere than those at wavelengths greater than 1.5 - 2 cm. The worst culprit is water vapor because it is not uniformly distributed. A handsome cumulus cloud can produce a fluctuation more than four orders of magnitude greater than the ΔT we seek to measure. Beam switching can help to reduce this source of noise, but unless one makes observations in good weather, atmospheric noise can completely dominate system noise. Hence the words of Parijskij in his 1973b article: "...we have used only recordings obtained in clear, settled, frosty weather". The presence of strong water vapor lines in the earth's atmosphere also makes it impossible to observe at wavelengths $\lesssim 1.5$ cm except in narrow wavelength "windows". The "windows" available are those at $\lambda \simeq 9$ mm and $\lambda \simeq 3$ mm.

While system and atmospheric noise together fix lower limits on the useful range of λ, the upper limit is determined by a quite different consideration: emission from local sources, either our own Galaxy or extragalactic radio sources. On small angular scales, emission from our own Galaxy will probably not present insuperable problems. It is, however, the predominant source of systematic errors in measurements of the large-scale anisotropy of the cmb, and we shall return to this issue later. More pertinent to our present discussion is the contribution of extragalactic radio sources, such as radio galaxies and quasars. Even if we choose to observe in a region where there are no catalogued radio sources, very faint radio sources randomly distributed on the sky may mask or mimic the fluctuations in the cmb we seek. The vast majority of such radio sources have power law spectra $I(\nu) \propto \nu^{-\alpha}$, with $\alpha > 0$ (typically 0.7), whereas $I(\nu)$ of the cmb is $\propto \nu^2$ in the Rayleigh-Jeans region. Therefore the contribution of such local sources rapidly becomes dominant as the frequency of observation is lowered, or λ is raised (Longair and Sunyaev, 1969). This consideration will almost certainly limit observations of fluctuations in the cosmic microwave background to wavelengths < 10 cm. As receivers are improved, permitting more sensitive measurements, this source of error may come to be the dominant one. If so, we will be forced to move to shorter and shorter wavelengths for the reason set out above. Unfortunately, the properties of radio sources, such as their number per steradian and their spectral index α, are not as well known at higher frequencies. Longair and Sunyaev (1969) made this point clearly. Better surveys at centimeter wavelengths are now becoming availa-

ble (see Jauncey, 1977, and references therein) which may help us to deal better with the problem of noise introduced by "local" sources.

For the reasons set out above, published searches for fluctuations in the cmb have been carried out in the wavelength range $3.5\ \text{mm} \leqslant \lambda \leqslant 4\ \text{cm}$. Most have been drift scans. One observational program which did not employ this technique (Boynton and Partridge, 1973) encountered severe problems because of side lobe radiation.

8. Required Corrections

The data from a typical search for temperature fluctuations might consist of 10–50 measurements of the intensity of the cmb at different points on the sky. A representative plot is shown in figure 3.

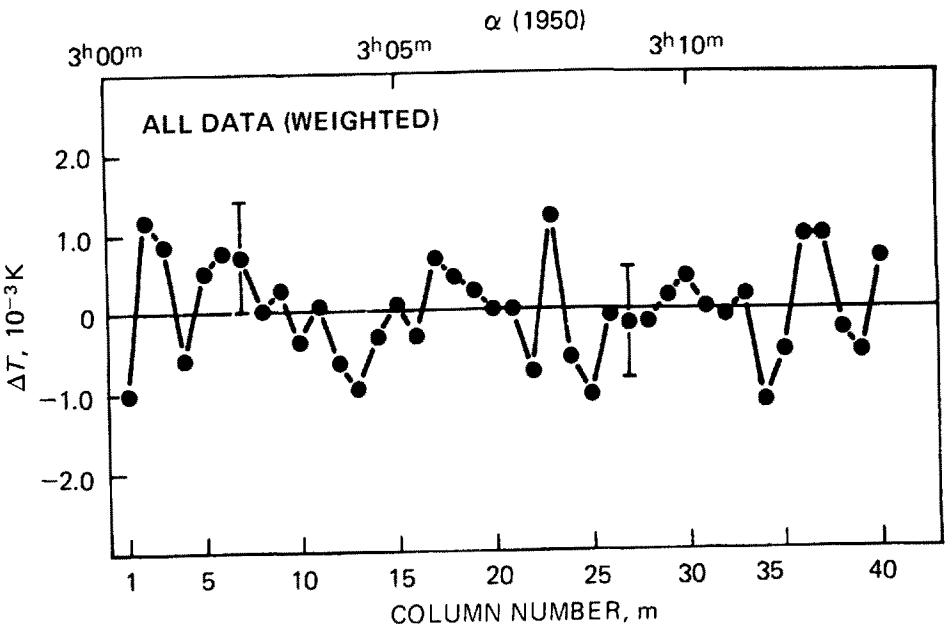

Fig. 3. Results of a drift scan made at $\lambda = 9$ mm, with an antenna beam of ~ 3.6 arcminutes width. The length of the scan was ~ 40 arcminutes. Typical error bars are shown. These data are preliminary results; the final results will appear in Partridge (1979).

We need next to examine what such a plot can tell us about the amplitude and average scale of possible anisotropies in the microwave background. Two major steps have to be taken. First, the raw numbers

must be corrected for instrumental inefficiency and absorption by the earth's atmosphere, and then converted to true thermodynamic temperatures.* The first of these corrections is a matter of judgment. The efficiency of a radio telescope, for instance, depends on whether one is seeking temperature fluctuations small compared to the main beam solid angle or fluctuations which essentially fill it. In the latter case, the quantity ε_b in equation 1 will be larger. This question is discussed further elsewhere (Partridge, 1979). Finally, if the observations are made in the form of drift scans, account must be taken of the fact that a source moves relative to the telescope, and therefore moves through the main beam solid angle. This in turn results in a "smearing" and slight loss in sensitivity. Again, this matter is further discussed elsewhere (Partridge, 1979). I have mentioned each of these corrections separately because not all of them have been made to the data which appear in the various published searches for fluctuations in the cmb.

9. Statistical Analysis

Once the data have been corrected one can try to determine whether statistically significant fluctuations in the sky temperature have been seen. A first approximation is simply to look at the data - clearly the fluctuations in ΔT shown in figure 3 lie at or below 1.2×10^{-3} K = 1.2 mK. A slightly more quantitative assessment is to compute the standard deviation of the data points.

These rough methods ignore the contribution of receiver noise to the scatter of the data points. Indeed, looking at figure 3 tells us nothing about fluctuations on the sky except that their rms value must be less than or equal to 1.2 mK. Conceivably all the variation could be due purely to instrumental noise. Conklin and Bracewell (1967) were the first to point out that there exist ways to subtract out system noise, at least on a statistical basis. This technique has been refined and employed by Parijskij in the analysis of his data

---

*Up to this point in this article, I have followed the usual radio astronomical convention of employing antenna temperature, defined by

$$T \text{ (antenna)} = \frac{SA}{2k} ,$$

where S is flux density (see Kraus, 1966, for instance). Except in the Rayleigh-Jeans region of the cmb spectrum, values of $\Delta T/T$ measured in antenna temperature are different for values of $\Delta T/T$ expressed in true, thermodynamic, temperature (see Boughn, Fram and Partridge, 1971). The required correction is small for all of the measurements discussed below.

(Parijskij, 1973b, 1977). Essentially, one makes use of the fact that the receiver noise contribution to scatter in the data decreases as more and more observations of each point on the sky are accumulated. Ideally, if s_r is the rms noise observed in a single observation for a single point, the rms receiver noise for n observations should decrease as $\sigma_r = s_r/\sqrt{n}$ (from eqn. 1). The observations will contain scatter from both receiver noise and possibly real fluctuations in the sky:

$$\sigma_o^2(n) = \sigma_s^2 + \sigma_r^2 = \sigma_s^2 + \frac{s^2}{n} \qquad (2)$$

The intercept of a plot of σ_o^2 versus 1/n gives σ_s^2.

This method assumes that receiver noise will decrease as $t^{-1/2}$. While this may not in fact be so, it will certainly not decrease faster. Therefore, the method offers a reliable upper limit on σ_s^2.

What is the precision with which σ_s^2 can be found by this method? It has been claimed that the error can be determined from the scatter in the plot of $\sigma_o^2(n)$ vs. $\frac{1}{n}$. However, it seems to me that the individual points in the plot are not independent (all the data are used to determine each), and therefore any estimate of the error in the intercept must take this dependence into account. Failure to do so will underestimate the true error in the estimate of σ_s^2.

An alternative method, employing the same basic principle of removing receiver noise, was developed by John Deeter and is described in Boynton and Partridge (1973). Suppose one has a series of n drift scans, each covering m independent points on the sky. Arrange the data in a n by m matrix. Recall that the data represent small differences in intensity between two nearby regions in the sky if beam switching is employed. Therefore each element in the matrix can be written as

$$X = I_S + I_A + I_G - (I_S' + I_A' + I_G') + N \qquad (3)$$

I_S is the intensity of the cmb at one position of the beam, I_A is the intensity contributed by the earth's atmosphere, and I_G is ground radiation into the side lobes. The primed quantities refer to the other position of the beam. N represents the contribution of randomly variable receiver noise. If drift scans are employed, so that the telescope is fixed, $I_G = I_G'$ to high accuracy, and any small random fluctuations in the difference $(I_G - I_G')$ can be lumped in with N. The same applies to the atmospheric contribution $I_A - I_A'$, providing the beam switch is rapid. In this approximation

$$X = I_S - I_S' + N$$

and we note that the true sky variance is given by

$$\sigma_s^2 \equiv \frac{1}{2} \text{var}(I_s - I_s') = \frac{1}{2}(\text{var } X - \text{var } N) \qquad (4)$$

Since the noise represented by N is due mainly to the receiver, we can return to our earlier nomenclature and write var(N) in (4) as σ_r^2. Hence if we average the data over rows of the n by m matrix, that is if we average over the scans, we will obtain a column-to-column variance which contains both the sky variance and a variance component due mainly to receiver noise. A similar analysis, sampling row-to-row differences, would yield the variance in N, or σ_r^2, exclusive of any sky variation. A comparison of "column variance", $\sigma_r^2 + 2\,\sigma_s^2$ to "row variance", σ_r^2, can then be used to determine $\sqrt{\sigma_s^2}$, or to set upper limits on the sky fluctuations (Boynton and Partridge, 1973).

It remains only to determine the confidence level of the upper limit (or the error in the estimate of σ_s). For the case $\sigma_s^2 < \sigma_r^2$, that is, where receiver noise dominates any real sky fluctuations, we can use the test described in Boynton and Partridge (1973). We wish to test the hypothesis that $\sigma_s^2 = 0$ against the hypothesis that σ_s^2 is equal to some <u>non-zero</u> value $\hat{\sigma}^2$. We begin by constructing the statistic

$$\delta = \sum_m \frac{x_m^2}{\sigma_m^2(\sigma_m^2 + \hat{\sigma}^2)} \qquad (5)$$

Note that σ_m^2 are the row variances, and the $(\sigma_m^2 + \hat{\sigma}^2)$ are the column variances <u>if</u> $\hat{\sigma}^2 = 2\,\sigma_s^2$. Hence the terms

$$\frac{x_m^2}{(\sigma_m^2 + \hat{\sigma}^2)}$$

in (5) are independent unit-normal variables when $\hat{\sigma}^2 = 2\,\sigma_s^2$, so that the statistic δ would be a χ^2 variable except for the additional σ_m^2 factor in the denominator, which may be thought of as a weight. However, if we define

$$R \equiv \sum_m \sigma_m^{-2} / \sum_m (\sigma_m^{-2})^2$$

then the product $R\,\delta$ is approximately distributed as χ^2 on ν degrees of freedom, where

$$\nu = R \sum_m \frac{1}{\sigma_m^2}$$

The χ^2 test permits us to estimate a confidence level, that is to conclude that the hypothesis that the sky fluctuations are greater than some value can be excluded at a given confidence level. This

analysis has been applied to the data of Boynton and Partridge (1973) and Partridge (1979), to which those interested in the finer details are referred.

This method does not have the drawbacks inherent in the method discussed earlier. However, it is most accurate for testing null or very small sky variances. If the observations improve to the point that $\sigma_r^2 \ll \sigma_s^2$, new statistical methods may be required. Fortunately (or more truly, unfortunately) we do not yet have measurements so good that we need to be concerned about this problem.

10. A Review and Analysis of Existing Measurements

Having explained the techniques of observation and data analysis (including some of the pitfalls), I will now try to review the existing measurements. Table 1 contains what I believe to be a complete summary of the observational results. Some of the earlier work (by Conklin and Bracewell, 1967; Penzias, Schraml and Wilson, 1969; and Boynton and Partridge, 1973) has clearly been superseded, but of course represented much hard work with the receivers available five to ten years ago. The relevant measures are those of Carpenter at al (1973), Stankevich (1974), Parijskij and his colleagues (1973a, 1973b, 1977), and Caderni et al (1977), all of which have been published, and the unpublished works by J.C. Pigg and by me. Let me review each briefly.

In some cases, I have applied corrections which I believe to be necessary to the published results (see final column of Table 1). The retrospective application of corrections to other people's data is a risky business. Let me apologize in advance to my colleagues whose work I am about to review - especially if I appear to have misread their work.

Carpenter et al (1973) worked at $\lambda = 3.6$ cm, employing the 64 meter antenna at Goldstone, California. This combination of receiver and antenna gave them a beam of halfwidth $\sim 2'$. They have taken into account all of the relevant corrections described above.

For his measurement, Stankevich employed the 64-meter telescope in Parkes, Australia. Because of the long wavelength he used, $\lambda = 11$ cm, a substantial correction was required for the fluctuations produced by weak discrete sources in his beam. Indeed, as he points out, a reasonable extrapolation from a survey at 408 MHz suggests that all the fluctuation he sees "...could be attributable to unresolved discrete sources". In fact, interferometric observations at 11 cm, by Martin, Partridge and Rood, to be discussed below, and the work of

Observers	Wavelength cm.	Angular Scale	Reported or Published $\Delta T/T^*$	Corrected $\Delta T/T^*$
1. Conklin and Bracewell (1967)	2.8	10′	$\leqslant 1.8 \times 10^{-3}$	
2. Penzias et al (1969)	0.35	2′	$\leqslant 9 \times 10^{-3}$	$\lesssim 2 \times 10^{-2}$
3. Boynton and Partridge (1973)	0.35	$\sim 1.5'$	$\leqslant 3.7 \times 10^{-3}$	$\leqslant 2 \times 10^{-3}$
4. Carpenter et al (1973)	3.6	$2'-1^{\circ}$	$\leqslant 7 \times 10^{-4}$	
5a. Parijskij (1973a)	2.8	$3'-1^{\circ}$	$\leqslant 3 \times 10^{-5}$	$\leqslant 4 \times 10^{-4}$
5b. Parijskij (1973b)	4.0	$\sim 12' \times 40'$	5×10^{-5}	$\leqslant 1.6 \times 10^{-4}$
6. Stankevich (1974)	11.1	8′-20′	$\leqslant 1.5 \times 10^{-4}$	$\lesssim 3 \times 10^{-4}$
7. Caderni et al (1977)	0.13	30′	$\leqslant 1.2 \times 10^{-4}$	(?)
8. Partridge (1978)	0.9	$\sim 3'$	$\lesssim 1.5 \times 10^{-4}$	
9. Pigg (1978)	2.0	1.3′	$\sim 4 \times 10^{-4}$(?)	
10. Parijskij (1977)	4.0	5′-150′	$\leqslant 8 \times 10^{-5}$ to $\leqslant 1.3 \times 10^{-5}$	see text

*Assuming $T = 2.7^{\circ}K$; upper limits are generally quoted as 2σ or at the 95% confidence level.

Table 1. A list of all published and unpublished measurements of the small-scale anisotropy of the cmb. In some cases such as (3) and (5), the authors have revised their published values. In other cases, I have attempted to apply some of the corrections described in this article. These are discussed more fully in the text.

Wall and Cooke (1975), suggest that his pessimistic view may in fact be justified.

Stankevich's brief communication does not provide enough detail to reveal whether the corrections described above have been applied to his data. It is clear from the parameters cited in his article, however, that he did not take in account the efficiency of the Parkes telescope - the factor ($\varepsilon_a \varepsilon_b$). Including this factor alone degrades the sensitivity of his limit on sky temperature fluctuations by a factor of 2.

In 1973, Parijskij reported two upper limits on $\Delta T/T$. For the first (1973a) he used a 2.8 cm receiver, also on the 64-meter dish at Goldstone, Cslifornia. His original published upper limit of $\lesssim 3 \times 10^{-5}$ has been revised by him to $\leqslant 4 \times 10^{-4}$ (see Boynton, 1978). The second (1973b) measurement was at $\lambda = 4$ cm, using the Pulkovo radio telescope, which has a main beam solid angle of $\sim 12' \times 40'$. As was the case for Stankevich's result, his published upper limit needs to be corrected for telescope efficiency (I estimate $\varepsilon_a \varepsilon_b \sim .5$). I have also converted his value to a 2σ or ~95% confidence limit. These two corrections produce $\Delta T/T \leqslant 1.6 \times 10^{-4}$, which appears in the final column of Table 1. Finally, there is the question of the estimate of the error in σ_s^2 derived from the procedure of Conklin and Bracewell (1967) - see discussion above. In this particular paper, the error in the estimate of $\sqrt{\sigma_s^2}$ is given as 4×10^{-5} K.* However, from the data provided by the author, it is clear that the actual observed scatter, when all 12 recordings had been averaged together, was $\sim 4\text{-}5 \times 10^{-4}$ K*, more than an order of magnitude greater. I would have expected the error in the estimate of $\sqrt{\sigma_s^2}$ to be roughly comparable to the scatter $\sqrt{\sigma_o^2(12)} \approx 4\text{-}5 \times 10^{-4}$ K; hence I suspect Parijskij has underestimated the error in σ_s^2. This expectation is based on experience with my own data (Partridge, 1979), which I suppose to be receiver noise dominated like his. A precise resolution of the doubt I have raised would require access to Parijskij's data in more detail than appears in his article (1973b). Since this point is based on hunch rather than calculation, I have not taken it into account in the entry in the final column of Table 1.

Parijskij's 1977 paper with Petrov and Cherkov reports the results of a more refined search employing the RATAN-600 telescope at $\lambda = 4$ cm. Angular scales from 5' to 150' were searched; the published 2σ upper limits on $\Delta T/T$ range from 8×10^{-5} to 1.3×10^{-5}.

---

*Uncorrected for the effects listed early in this paragraph.

However, from a preliminary reading of the paper, it appears to me that the telescope efficiency ($\varepsilon_a \varepsilon_b$ in my notation) has not fully been taken into account. The stated figure for telescope efficiency, 87%, seems high. Again, I emphasize that the doubt I raise may be based on a misreading of their paper. In addition - perhaps because I have not yet seen an English version of the paper - I do not fully understand the data analysis. Using n = 20 (the number of scans) in their formula (1) seems to produce values of $\Delta T/T \gtrsim 3$ times larger than their published values.

The work of Caderni et al (1977) may represent the wave of the future in this field. They employed a wide-band bolometric detector as their receiver, and a small flux collecting antenna. Corrections for telescope efficiency and absorption by the atmosphere appear to have been included (since the instrument was calibrated using astronomical sources). It is not clear whether the correction for "smearing" was included, nor is much detail provided on the analysis of the data. I suspect any changes introduced by such considerations would not alter their final results very much. Note the relatively large angular scale of the measurements, imposed by experimental constraints discussed by the authors.

Pigg also worked with the 64-meter Goldstone antenna, at $\lambda = 2$ cm. His observations were made in a region of high galactic latitude. Preliminary analysis suggests $\simeq 1$ mK as an upper limit on the fluctuations in the cmb on a scale of 1.4′, his half-power beamwidth. I do not know how many points he observed, or which of the various corrections discussed above were applied to his data. Note that the angular scale for which these observations were made is the smallest of any of the sensitive measurements contained in Table 1.

My own work will be discussed in considerably more detail elsewhere (Partridge, 1979). Here, let me simply summarize the relevant observational details. Measurements were made at $\lambda = 9$ mm using the 11 meter telescope of the National Radio Astronomy Observatory in Tucson. Drift scans were made of a small strip of the sky at declination = 80°, right ascension = 3^{h}, and another smaller strip at the same declination, but right ascension = 9^{h}. The half-power beamwidth of this instrument was 3.6′. Once all of the corrections discussed above had been made, the following preliminary conclusions could be drawn: - At the 95% confidence level fluctuations in the sky temperature on angular scales substantially smaller than 3.6′ were smaller than 0.7 mK, that is, $\Delta T/T \lesssim 2.5 \times 10^{-4}$. The experiment permitted me to set somewhat more sensitive limits on fluctuations of approxi-

mately the same angular scale as the main beam solid angle, that is $\theta \sim 3\text{-}4'$. Here, the 95% confidence limit appears to be $\Delta T/T \lesssim 1.5 \times 10^{-4}$.

11. Interferometric Search for Fluctuations in the cmb

The angular scale of 1.4′ reached by Pigg is about the lower limit that can be achieved by using conventional, filled aperture, radio telescopes. Perhaps the best one could do is the 100 meter telescope in Bonn, which, if used with a 1.25 cm receiver, would give a half-power beamwidth of ~40". To search on smaller angular scales, which may be of interest if we hope to see protogalactic fluctuations, one has to resort to new methods. One possibility is interferometry. Using a spaced array of radio telescopes as an interferometer offers far greater angular resolution, and the additional advantage that one can obtain a two-dimensional map of the sky.* In principle, then, interferometry offers the ideal approach to a search for small scale fluctuations in the cmb. Unfortunately, one pays an extremely heavy price in sensitivity. In crude terms the difficulty is that conventional interferometers look at too many individual points in the sky. Thus the effective amount of integration time spent on each independent point on the sky is small. Nevertheless, observations of this sort have been undertaken. The first experiment of this nature that I know of was performed several years ago, for a different purpose, by Goldstein, Marscher and Rood (1976). Their measurements were made at $\lambda = 21$ cm, where the contribution from faint radio sources no doubt overwhelmed fluctuations in the cmb. In addition, their experiment was not particularly sensitive, setting limits on ΔT of $\lesssim 90$ mK.

Martin, Rood and I have attempted to repeat and refine this experiment. We worked with the same interferometer, the three element array in Green Bank, West Virginia, but we used shorter wavelengths, 3.7 cm and 11 cm. We observed a region centered at $\Delta = 80^{\circ}08'$, $\alpha = 3^{h}10^{m}$ using baselines of 1900, 1800, 1200, 600, and 100 meters. Because of the high declination of the source, and the reasonably large number of baselines employed, the synthesized beam of the interferometer was well defined or "clean". The beam is shown in figure 4. We obtained a total of eight nights of observation. The interferometric map of the region, with all data included, is shown in figure 5. This map was then analyzed using a variant of the method employed by

*This would be a particular advantage if the fluctuations sought were not highly symmetrical.

Goldstein et al (1976).

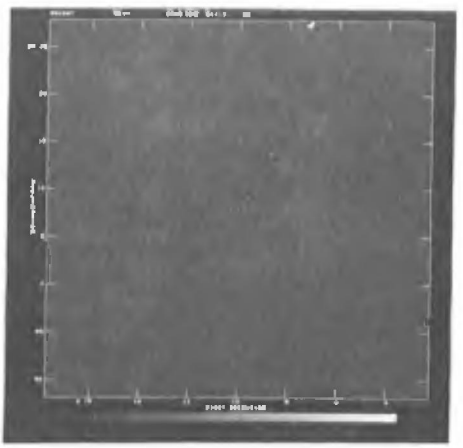

Fig. 4. The synthesized interferometer beam used by Martin, Partridge and Rood to search for small angular scale fluctuations. The "rings" in the interference pattern result from the small number of baselines used

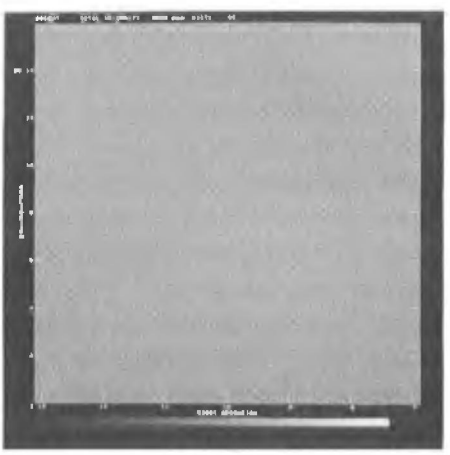

Fig. 5. A map of the sky at $\lambda = 3.7$ cm obtained using the Green Bank interferometer

This method allows us to make an approximate correction for the fluctuations introduced by receiver noise. The analysis of our data is not yet complete, but preliminary values for the upper limits on fluctuations in the sky are:

$$\text{at 11 cm, } \sqrt{\sigma_s^2} \leqslant 0.45 \pm 0.22 \times 10^{-3} \text{ Jansky, and}$$

$$\text{at 3.7 cm, } \sqrt{\sigma_s^2} \leqslant 0.18 \pm 0.12 \times 10^{-3} \text{ Jansky.}$$

The conversion from flux units to sky temperature presents more problems for an interferometric search than for a conventional radio search. I prefer to skip over these details here, since they will be discussed fully in the paper which Martin, Rood and I are preparing. At the moment, we are not entirely convinced that the conversion method used by Goldstein et al is correct. But if we do employ this method to convert the results of our own experiment we find the following preliminary results:

$$\Delta T \leqslant 18 \pm 9 \text{ mK at 11 cm,}$$

$$\Delta T \leqslant 7 \pm 5 \text{ mK at 3.7 cm.}$$

The price paid in sensitivity for the higher resolution afforded by interferometry is quite clear. These upper limits to fluctuations in the sky temperature apply to angular scales smaller than the synthesized beam. In the case of the 11 cm observations, this scale is ~12"; for the shorter wavelength measurements, it is ~4".

Although this first attempt did not place interesting limits on fluctuations in the cmb, some useful results did emerge. In effect, the experiment represented an extension of the source surveys at 11 cm and 4 cm to somewhat fainter flux densities. For instance, at 11 cm the measured fluctuation level was 0.45 ± 0.22 mJy. From the way we constructed our interferometer maps, we know that this was the mean level of fluctuation detected in cells of solid angle $5' \times 5'$. If we now accept the argument of Longair and Sunyaev (1969) that most of the fluctuation due to discrete sources arises from the presence of the single brightest source in the solid angle under study, then our measurements suggest the presence of one source of ~ 0.45 mJy in each $5' \times 5'$ area of the sky (equivalent to 2×10^{-6} ster). This result is in good agreement with an extrapolation of the survey of Wall and Cooke (1975). The work at $\lambda = 3.7$ cm may be used to corroborate the extrapolation made by Longair and Sunyaev (1969); our measured values are in good agreement with their predictions down to a level of 2-3 mJy. We conclude that for any filled aperture telescope now extant, fluctu-

ations produced by unresolved radio sources will swamp fluctuations in the cmb if observations are made at 11 cm; but will not at 4 cm.

12. The Large Scale Isotropy of the Universe

Two of the basic assumptions underlying the Friedman models of cosmology are that the Universe is homogeneous, and that it is isotropic. Included in the latter is the requirement of zero shear. As we have seen, the Universe is manifestly not homogeneous. To date, however, we have no evidence for the existence of systems of galactic masses or larger for which $GM/Rc^2 \simeq 1$. Hence on the large scale the assumption of homogeneity is approximately valid.

As has been pointed out elsewhere in this volume, the cmb provides the strongest observational evidence supporting the validity of the second assumption of Friedman cosmologies as well. The observed isotropy of the cosmic microwave background sets stringent limits on anisotropic cosmological expansion and shear.

Several sorts of large angular scale anisotropy may be present in the cmb. A dipole anisotropy may be produced by the Doppler effect if the observer is in motion relative to the cmb. Our observations of the cmb are not yet sensitive enough to detect the velocity of the Earth. Measurements of the dipole anisotropy yield only a value for the solar velocity. Since the velocity of the sun in its orbit around the center of the Galaxy is known to reasonable accuracy, one can correct the observed velocity to obtain the speed and direction of the motion of the center of our Galaxy. This number is of interest in its own right, and the value may also be used to set limits on certain classes of anisotropic cosmological models (Thorne, 1967; Hawking, 1969).

Quadrupole anisotropy is produced by anisotropic expansion. This topic is discussed more fully by McCallum in this volume. One frequently overlooked fact deserves further mention, however. Observers seem frequently to have forgotten that a simple quadrupole form will be present only for certain classes of anisotropic cosmological models. As Novikov and others (1968, 1977) have been at pains to point out, the general situation is more complicated, especially in open ($\rho_0 < \rho_c$) models. In particular, for open anisotropic models, large departures from the mean background temperature may occur only in a rather limited solid angle, given approximately by $\Omega = (\rho_0/\rho_c)^2$. A survey covering a limited part of the sky cannot reliably be used to rule out certain classes of anisotropic cosmological models.

With the exception of this last point, all these matters are considered more carefully by other contributors to this volume. Therefore, let me turn immediately to the present status of the observations. To begin with, let me emphasize that I have not been directly involved in measurements of this sort (except for an abortive attempt in 1976) for the last ten years. Therefore, all the work I report in this section will be other workers', and I would urge you to consult their papers for further details.

In Table 2 I have summarized the results of recent and reasonably accurate measurements of the dipole anisotropy of the cmb. Some earlier work has been omitted. The values ascribed to Corey and Wilkinson and to the Berkeley group (Smoot et al, 1977) are preliminary. Both groups (and an additional group at MIT) are pursuing these investigations. Nevertheless, the data already in hand do agree moderately well on a velocity of $v_s \sim 300$ km/sec in the direction $\delta \sim 0^o$, $\alpha = 12^h$. While the speed is about equal to the speed of the sun in its circular path about the center of the Galaxy, the direction is quite different. Therefore, when v_s is corrected for solar motion in the Galaxy, we obtain a larger value for the speed of the Galaxy as a whole. The velocity of the Galaxy appears to be $v_G \simeq 600$ km/sec towards $b = 30^o$, $\ell = 260^o$, in Galactic coordinates. It is worth noting that the value of v_G obtained from measurements of the cosmic microwave background is not in good agreement with the same quantity derived from measurements of the redshifts of nearby spiral galaxies (Rubin et al, 1976). To reconcile these two sets of observations, one has to assume motion of the entire swarm of galaxies within ~ 100 Mpc relative to the cmb.

All the measurements of the large-scale isotropy of the cmb have been made in the Northern hemisphere, so that a large fraction of the Southern sky has not been surveyed. As a consequence, it is not easy to separate out possible dipole and quadrupole moments* Nor, of course, can we rule out the possibility of more complicated angular structure, such as that suggested by Novikov. Observations in the Southern hemisphere are being planned. Until they have been completed, we can say with reasonable certainty only that the quadrupole component of the anisotropy is less than or about 2 mK, so that $\Delta T/T < 7 \times 10^{-4}$.

All of the good measurements of large-scale anisotropy have been carried out above all or most of the earth's atmosphere. This fact

---

*The determination of the declination component of the dipole anisotropy is also difficult.

DIPOLE

Observers	λ, cm	One declination* or whole sky?	ΔT, mK, 1σ error	α of max	δ of max (if deter.)
Partridge and Wilkinson (1967)	3.2	$\delta = -8^{\circ}$	2.2 ± 1.8	17^{h}	
Henry (1971)	3.0	w.s.	3.2 ± 0.8	$10^{h}.5 \pm 4^{h}$	$-30^{\circ} \pm 25^{\circ}$
Conklin (1972)	3.8	$\delta = +32^{\circ}$	2.3 ± 0.9	11^{h}	
Corey and Wilkinson (1976) (revised 1978)	1.5	w.s.	2.9 ± 0.7	$12^{h}.3 \pm 1^{h}.4$	$-21^{\circ} \pm 21^{\circ}$
Smoot et al (1977)	0.9	w.s.	3.2 ± 0.5	$11^{h}.3 \pm 0^{h}.4$	$+14^{\circ} \pm 7^{\circ}$

QUADRUPOLE

Partridge and Wilkinson (1967)	3.2	$\delta = -8^{\circ}$	2.7 ± 1.9	7^{h}, 19^{h}	
Conklin (1972)	3.8	$\delta = +32^{\circ}$	1.4 ± 0.8	6^{h}, 18^{h}	
Smoot et al (1977)	0.9	w.s.	<1.0	?	

*Early observations were scans made at a fixed δ ; later observations scanned much of the visible northern sky.

Table 2. Measurements of the large-scale anisotropy of the cmb

reveals the fundamental dilemma facing observers in this field. Observations at short wavelength are plagued by emission from the earth's atmosphere and the insensitivity of high frequency receivers discussed earlier. On the other hand, observations at long wavelength run into the serious systematic problem of anisotropic radio emission from our own Galaxy. A decade of experience has shown that it is not possible to make measurements of sufficient accuracy from the surface of the earth. Hence observers in this field have resorted to balloon and aircraft flights to obtain their measurements, and these efforts are continuing. The ultimate experiment may be the Cosmic Background Explorer, a small satellite equipped with sensitive radiometers at several different frequencies above 10 GHz. In space, where the Earth's atmosphere is no longer a problem, the short wavelength limit for accurate observations will probably be set by thermal emission from dust in our Galaxy. This material radiates dilute blackbody (or greybody) radiation with a characteristic temperature of ~ 100 K. While, like the radio emission which sets the upper wavelength limit for these observations, thermal emission from dust is largery confined to the plane of the Galaxy; there is some evidence that the concentration of dust and hence its emission is inhomogeneous on angular scales of a few degrees (Weiss, private communication). If so, attempting to subtract out the contribution of thermal dust emission will be made more difficult.

These problems lie in the future. Even now, however, measurements of the cmb have revealed the extremely important cosmological result that our Galaxy is in relatively rapid motion with respect to the cosmic background. Even if the value of v_G changes slightly as more accurate observations are obtained, the major result is the very magnitude of v_G.

13. New Directions

Here and there in this article, I have mentioned matters that need further study or possible future observations. Since what I have said has been directed at a group of young scientists interested in cosmology, let me end by speculating briefly on what the next few years will bring in this field.

Within the next few years, we should have more accurate observational results for the large-scale anisotropy. Possible sources of systematic error in the large-scale measurements, such as radio and thermal dust emission from our Galaxy, will be better understood. Meas-

urements made in the Southern hemisphere should help strengthen the limit on quadrupole anisotropy. The one remaining observational prize would be the detection of the kind of isolated cosmological anisotropy suggested by Grishchuk et al (1969) and Novikov (1977). Measurements of the dipole and quadrupole moments, now or soon to be available, should permit us to restrict the solid angle in which this complicated cosmological anisotropy signal might be found. A careful search should then be carried out at high enough angular resolution to resolve the complicated structure of this anisotropy (angular scale $\theta \simeq \sqrt{\rho_0/\rho_c}$). The detection of such a characteristic anisotropy would provide direct evidence that the Universe is open (low density).

There is also a good deal of work to be done in searching for anisotropies on smaller angular scales. Conventional radio astronomical measurements could lower limits on $\Delta T/T$ on angular scales of a few arcminutes by perhaps another factor of 3. The limits in this range of angular scale are already the best we have. I would like to see that effort pursued. I also think it would be useful to improve our limits on $\Delta T/T$ on somewhat larger angular scales, those corresponding to fluctuations which were not inside the light horizon at the epoch of recombination. As Weinberg (1972) has pointed out, such large mass aggregates were not causally connected, so we have no a priori reason to expect them to be homogeneous, at least in conventional cosmological theory. My hunch is that no substantial anisotropy on scales of 2°-3° will be found, but if it is much of our conventional thinking about cosmology might have to be revised. I plan to look into this matter, in collaboration with colleagues at the Universities of Tromsö and Bologna, this next year.

For all angular scales from a few arcminutes up to several degrees, conventional radio astronomical techniques may soon be superseded by bolometric measurements (see Boynton, 1978). Although in radio astronomical terms, the system noise of bolometers is far larger than for conventional radio receivers, bolometers offer the advantage of huge bandwidth, $\Delta\nu$. One important bolometric measurement of small-scale anisotropy has already been reported (Caderni et al, 1977), and others are being planned. Bolometric measurements will necessarily have to be made at high frequency (to obtain the desired bandwidth), and hence will encounter problems with the Earth's atmosphere. Hence the searches may have to be carried out from aircraft or balloons, or carried out only in moments of exceptional atmospheric stability. Only new techniques of this sort, I would guess, will permit us to reach sensitivities well below the limit of $\Delta T/T \sim 10^{-4}$.

Finally interferometric studies in this field are only just beginning. The ability to map anisotropies, so as to reveal their characteristic shape, would be a real advantage. A new generation of interferometers such as the Very Large Array in New Mexico, may eventually permit us to reach fluctuation levels of $\sim$1 mK.

Acknowledgements

This work in its final form owes much to the comments of my colleagues in Jodłowy Dwór. I would also like to acknowledge the very considerable help of a Haverford student, Michael Gregg, in organizing this manuscript. The preparation of this report, and some of my research reported in it, was supported by a grant from the U.S. National Science Foundation, and also by the Faculty Research Fund of Haverford College.

References

Since other authors in this volume have dealt with the theory, in more detail, I have not attempted to make this list of references complete for the theoretical work. Only representative citations are given. The reference list is much more complete for the observational work, on which I have put the most emphasis.

Anile, A.M., Danese, L., De Zotti, G., and Motta, S. 1976 Ap. J. Letters, 205, L59.
Barrow, J. 1976, Month. Not. Roy. Ast. Soc., 175, 359. See also Month. Not. Roy. Ast. Soc., 81, 719.
Boughn, S.P., Fram, D.M., and Partridge, R.B. 1971, Ap. J., 165, 439.
Boynton, P.E., and Partridge, R.B. 1973, Ap. J., 181, 243.
Boynton, P.E. 1978 in I.A.U. Symposium 79, ed. M.S. Longair (Reidel, Dordrecht).
Caderni, N., Fabbri, R., DeCosimo, V., Melchiorri, B., Melchiorri, F., and Natale, V. 1977, Phys. Rev., D 16, 2424.
Carpenter, R.L., Gulkis, S., and Sato, T. 1973, Ap. J. Letters, 182, L61.
Conklin, E.K., and Bracewell, R.N. 1967, Nature, 216, 777.
Corey, B.E., and Wilkinson, D.T. 1979, to be published. See also Bull. A.A.S., 8,351, 1976.
Dautcourt,G. 1969, Astrophys. Letters, 3, 15. See also Astron. Nachrichten, 298, 141, 1977.
Davis, M. and Wilkinson, D.T. 1974, Ap. J., 192, 251.
Doroshkevich, A.G., Zel'dovich, Ya.B., and Sunyaev, R.A. 1977a, Ast. Zh., in press.
Doroshkevich, A.G., Novikov, I.D., and Polnarev, A.G. 1977b, Ast. Zh., 54, 932. English version in Sov. A. J., 21, 529.
Fabian, A.C., and Rees, M.J. 1979, preprint.
Field, G.B. 1975, in Stars and Stellar Systems, Vol IX, ed. A. Sandage, M. Sandage and J. Kristian (Univ. of Chicago Press, Chicago), p. 359.

Goldstein, S.J., Marscher, A.P., and Rood, R.T. 1976, Ap. J., 210, 321.
Gott, J.R., Gunn, J.E., Schramm, D.N., and Tinsley, B.M. 1974, Ap. J., 194, 543.
Gott, J.R., and Rees, M.J. 1975, Astron. and Astrophys., 45, 365.
Grishchuk, L.P., Doroshkevich, A.G., and Novikov, I.D. 1969, Sov. Physics JETP, 28, 1210.
Gunn, J.E., and Peterson, B.A. 1965, Ap. J., 142, 1633.
Hawking, S.W. 1969, Month. Not. Roy. Ast. Soc., 142, 129.
Henry, P.S. 1971, Nature, 231, 516.
Jauncey, D.L. 1977, I.A.U. Symposium 74 (Reidel, Dordrecht).
Jeans, J.H. 1928, Astronomy and Cosmogony (Cambridge Univ. Press, Cambridge).
Jones, B.T. 1976, Rev. Mod. Phys., 48, 107.
Kraus, J.D. 1966, Radio Astronomy (McGraw-Hill Book Co., Inc., New York).
Larson, R.B. 1974, Month. Notices Roy. Ast. Soc., 166, 585.
Larson, R.B. 1977 in The Evolution of Galaxies and Stellar Populations, ed. B.M. Tinsley and R.B. Larson (Yale Univ. Press, New Haven), p. 97.
Lifschitz, E. 1946, Journal Phys. USSR, 10, 116.
Longair, M.S., and Sunyaev, R.A. 1969, Nature, 223, 721.
Novikov, I.D. 1968, Ast. Zh., 45, 538. English version in Sov. A.J., 12, 427.
Novikov, I.D., 1977 in I.A.U. Symposium 74, ed. D.L. Jauncey (Reidel, Dordrecht), p. 335.
Ozernoi, L.M., and Chernin, A.D. 1967, Astr. Zh., 44, 1131. English version in Sov. A. J., 11, 907, 1968.
Ozernoi, L.M., and Chibisov, G.V. 1971, Ast. Zh., 47, 769. English version in Sov. A. J., 14, 615.
Ozernoi, L.M. 1974, in I.A.U. Symposium 63, ed. M.S. Longair (Dordrecht: Reidel), p. 227.
Parijskij, Yu.N. 1973a, Ap. J. Letters, 180, L47.
Parijskij, Yu. 1973b, Ast. Zh., 50, 453. English version in Sov. A. J., 17, 291.
Parijskij, Yu.N., Petrov, Z.E., and Cherkov, L.N. 1977, Pis'ma Ast. Zh., 3, 483. English version in Sov. A. J. Letters, 3, 263.
Partridge, R.B. 1974, Ap. J., 192, 241.
Partridge, R.B. 1975, in Proceedings of the First Marcel Grossman Meeting on General Relativity (North Holland).
Partridge, R.B. 1979, submitted to Ap. J.
Peebles, P.J.E. 1968, Ap. J., 153, 1. See also Peebles' Physical Cosmology.
Peebles, P.J.E. 1971, Physical Cosmology (Princeton Univ. Press, Princeton, N.J.).
Penzias, A.A., and Wilson, R.W. 1965, Ap. J., 142, 419.
Penzias, A.A., Schraml, J., and Wilson, R.W. 1969, Ap. J. Letters, 157, L49.
Press, W.H. and Schechter, P. 1974, Ap. J., 187, 425.
Rees 1971, in Proceedings of the International School of Physics Enrico Fermi, Course 47, ed. R.K. Sachs (Academic Press, New York), p. 315.
Rees, M.J. 1977, in The Evolution of Galaxies and Stellar Populations, ed. B.M. Tinsley and R.B. Larson (Yale Univ. Press, New Haven), p. 339.
Rees, M.J., and Ostriker, J.P. 1977, Month. Notices Roy. Ast. Soc., 179, 541.
Rubin, V.C., Thonnard, N., Ford, W.K., and Roberts, M.S. 1976, A. J., 81, 719.
Seldner, M. and Peebles, P.J.E. 1977, Ap. J., 215, 703, and references therein.

Shklovskii, I.S. 1960, Cosmic Radio Waves (Harvard University Press, Cambridge, Mass.).
Silk, J. 1968, Ap. J., 151, 459.
Silk, J. 1974, in I.A.U. Symposium 63, ed. M.S. Longair (Reidel, Dordrecht), p. 175.
Silk, J., and Wilson, M.L. 1979, preprint.
Smoot, G.F., Gorenstein, M.V., and Muller, R.A. 1977, Phys. Rev. Letters, 39, 898.
Stankevich, K.S. 1974, Ast. Zh., 51, 216. English version in Sov. A. J., 18, 126.
Sunyaev, R.A. and Zel'dovich, Ya.B. 1970, Astrophys. and Space Sci., 6, 358.
Sunyaev, R.A. 1971, Astron. and Astrophys., 12, 190.
Sunyaev, R.A. 1977a, in I.A.U. Symposium 74, ed. D.L. Jauncey (Reidel, Dordrecht), p.327.
Sunyaev, R.A. 1977b, Pis'ma Astron. Zh., 3, 491. English version in Sov. A. J. Letters, 3, 268.
Sunyaev, R.A. 1978, in I.A.U. Symposium 79, ed. M.S. Longair (Reidel, Dordrecht), p.393.
Sunyaev, R.A., and Zel'dovich, Ya.B. 1972, Comments on Space Sci. and Astrophys., 4, 173.
Thorne, K.S. 1967, Ap. J., 148, 51.
Tinsley, B.M., and Larson, R.B. 1977, The Evolution of Galaxies and Stellar Populations (Yale Univ. Press, New Haven).
Wagoner, R.V. 1973, Ap. J., 179, 343.
Wall, J.V., and Cooke, D.J. 1975, Month. Not. Roy. Ast. Soc., 171, 9.
Weinberg, S. 1972, Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity (John Wiley and Sons, Inc., New York).
Zel'dovich, Ya.B., and Sunyaev, R.A. 1969, Astrophys. and Space Sci., 4, 285.
Zel'dovich, Ya.B. 1978, I.A.U. Symposium 79, ed. M.S. Longair (Reidel, Dordrecht).
Zentsova, A.S., and Chernin, A.D. 1977, Pis'ma Astron. Zh., 3, 488. English version in Sov. A. J. Letters, 3, 266.

CONSTRAINTS ON THE POSSIBLE DISTORTIONS OF THE COSMIC BACKGROUND RADIATION SPECTRUM

G. De Zotti
Unità di Ricerca di Asiago-Padova del C.N.R.

1. Introduction

Professor Zeldovich (this volume) has devoted a few lectures to the distortions of the cosmic background radiation (CBR) spectrum which can be expected on theoretical grounds and has shown how much it could be learned from them on how the universe has evolved. I will present here the main results of an analysis of the presently available observational data, carried out by Dr. Danese and myself, aimed to investigate whether any evidence of deviations from a blackbody (BB) shape can already be found and to estimate the allowed ranges for the parameters defining the amplitude of the distortions (for a fuller account see Danese and De Zotti 1978).

2. Expected distortions

The following short description of the distorted spectra is based on the work by Zeldovich, Sunyaev and Illarionov (Zeldovich and Sunyaev 1969, Sunyaev and Zeldovich 1970, Illarionov and Sunyaev 1974a,b; for a more detailed review see Danese and De Zotti 1977).

An energy release occurring at a redshift z_h larger than

$$z_a \simeq 4\times10^4\hat{\Omega}^{-1/2}, \qquad \hat{\Omega} = (H_o/50)^2\Omega \quad (\Omega = \text{density parameter}) \tag{1}$$

leads to the formation of a Bose Einstein-like spectrum with a frequency dependent chemical potential μ which approaches a constant value μ_o in the spectral region where the free-free processes are negligible, i.e. at frequencies much larger than the dimensionless frequency

$$x_{CB} = h\nu_{CB}/kT_e \simeq 1.2\times10^{-2}\hat{\Omega}^{-7/8}\left[6.4 - \ln(10\hat{\Omega})\right]^{1/2} \tag{2}$$

where T_e is the electron temperature. At frequencies $\ll \nu_{CB}$, on the other hand, the free-free processes dominate and $\mu \to 0$. Fig. 1 illustrates the general behaviour of the equivalent thermodynamic temperature T_{eq} as a function of frequency. This distorted spectrum is

characterized by a marked drop in T_{eq} of amplitude

$$\left.\frac{\Delta T}{T}\right|_{max} \simeq 4.5\,\mu_o\,\hat{\Omega}^{-7/8} \qquad (\mu_o < x_{CB}) \tag{3}$$

occurring at a wavelength

$$\lambda_m \simeq 3.5\,\hat{\Omega}^{-7/8}\ \text{cm} \tag{4}$$

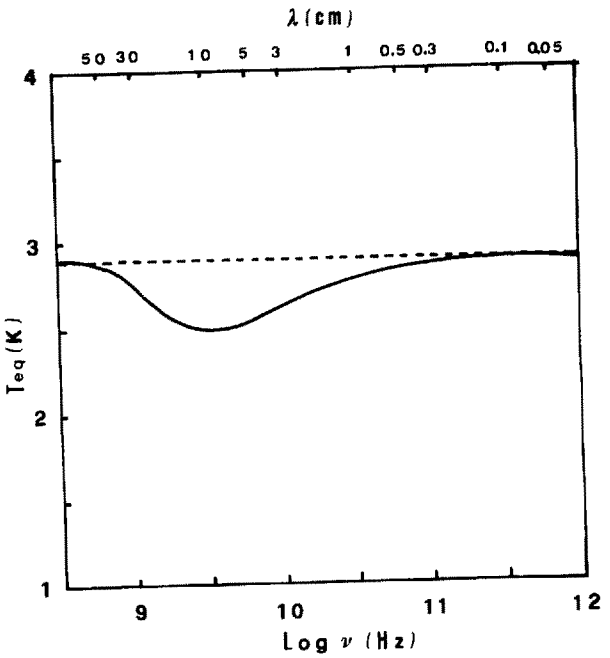

Fig. 1. T_{eq} as a function of frequency in the case of a BE-like distorted spectrum with $\hat{\Omega} = 0.3$, $\mu_o = 0.01$ and $T_R = 2.9°K$

The chemical potential μ_o is related to the fractional amount of energy released $\Delta\varepsilon/\varepsilon$ by

$$\mu_o \lesssim 1.4\ \Delta\varepsilon/\varepsilon \qquad (\mu_o \ll 1) \tag{5}$$

where the $<$ sign accounts for the fact that distortions originated at very early epochs ($z_h > z_\mu \cong 3.6\times10^5\,\hat{\Omega}^{-6/5}$) are smoothed out by the combined action of free-free and Compton processes.

If the energy release occurs at $z_h < z_a$ there is not enough time for the formation of a BE spectrum and the resulting distortion is, to some extent, peculiar to the process of energy dissipation which

was operating. However, on the limit $\Delta\varepsilon/\varepsilon \ll 1$, and if bremsstrahlung can be neglected, all spectra converge to the first order approximation

$$\eta_c \simeq \left[\exp(x) - 1\right]^{-1} \left[1 + \frac{u\, x \exp(x)}{\exp(x) - 1}\left(\frac{x}{\tanh(x/2)} - 4\right)\right] \qquad (6)$$

where η_c is the photon distribution function, $x = h\nu/kT_R$, T_R is the unperturbed radiation temperature and $u \simeq \Delta\varepsilon/4\varepsilon$. If the free-free processes are taken into account the photon distribution function writes:

$$\eta = \left\{1 - \left[1 - \eta_c\,(\exp(x_e) - 1)\right] \exp(-y_B)\right\} \left[\exp(x_e) - 1\right]^{-1} \qquad (7)$$

where η_c is the Compton distorted spectrum (eq. (6)), $x_e = xT/T_e$ and $y_B(x_e)$ is the free-free optical depth of the Universe. In the Rayleigh-Jeans (RJ) region, but above the frequency ν_B of free-free self absorption, the spectrum shows a frequency independent diminution of T_{eq} ($T_{eq} \simeq T_R(1 - 2u)$; see Fig. 2).

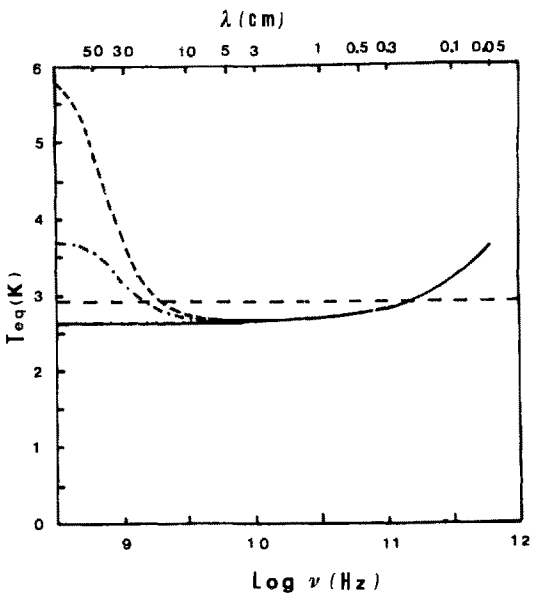

Fig. 2. T_{eq} as a function of frequency in the case of "late" energy release ($z_h \ll z_a$). The continuous line refer only to Compton distortions (eq. (6)). The dot-dashed and the dashed curves include the effect of bremsstrahlung (eq. (7)) in these cases: T_e is equal to its equilibrium value (eq. (8)) and $T_e = 2T_R$, respectively; also it has been assumed that $z_h = 2000$, $\Omega = 1$ and $T_R = 2.9°K$

At higher frequencies T_{eq} increases becoming equal to T_R at $\lambda \simeq 1 \div 2$ mm (depending on the actual value of T_R). Below ν_B, T_{eq} approaches the electron temperature T_e. The equilibrium value of T_e in the perturbed radiation field (7) is

$$T_e \simeq T_R(1 + 5.4\ u) \qquad (8)$$

Of course, in the case of direct electron heating T_e can be larger than its equilibrium value.

3. Comparison of the theoretical spectra with the data

The presently available observational data are discussed in Dr. Partridge's lectures (this volume). Let me only mention that one should be somewhat cautious in dealing with the submillimeter balloon measurements. In fact, even if the results of the Queen Mary College group (Robson et al. 1974) and of the Berkeley group (Woody et al. 1975) seem, at the first glance, to be in good agreement, there is a considerable discrepancy between the two groups as to the atmospheric parameters to be used to compute these results (see e.g. Clegg 1977).

In order to take into account all the possibilities we decided to carry out our analysis considering, in turn, only the ground based measurements (GB set), the GB plus the QMC data (GB + QM set), the GB plus the Berkeley data (GB + B set) and all the data together. The "best" values for the parameters and their ranges permitted by the data have been obtained by means of the usual "minimum χ^2" method (see e.g. Avni 1976, Cash 1976).

Table 1 gives the "best" values of the temperature and their 1σ uncertainties in the case of a BB spectrum.

Table 1. Results for a blackbody spectrum

Set of data	T (K)	$\chi^2_{\nu\,min}$
GB	2.73 ± 0.05	0.50
GB + QM	2.81 ± 0.04	0.74
GB + B	2.94 ± 0.03	2.25
GB + QM + B	2.94 ± 0.024	1.77

For the GB measurements $\chi^2_{\nu\,min}$ (= minimum χ^2 divided by the number of degrees of freedom) is quite small as a consequence of the fact in

some cases the published errors are not standard deviations, but are somewhat overestimated. It can be noticed that T_R increases when the submillimeter data are included, a fact that could be interpreted as an indication of some kind of distortions.

Tables 2 and 3 show that, indeed, if we trust the Berkeley data, there is an evidence for $\mu_o \neq 0$ at a confidence level $< 3\sigma$ and for $u \neq 0$ at a level $> 3\sigma$. The latter fact was first realized by Field

Table 2. Results for a BE-like spectrum

Set of data		"best" T_e (K)	"best" μ_o	3σ upper limit to μ_o	$\chi^2_{\nu min}$
GB	$\hat{\Omega}=1$	2.73	0	2.4×10^{-2}	0.53
	$\hat{\Omega}=0.1$	2.73	0	1×10^{-2}	0.53
	$\hat{\Omega}=0.03$	2.73	0	1×10^{-2}	0.53
GB+QM	$\hat{\Omega}=1$	2.84	9×10^{-3}	3.3×10^{-2}	0.74
	$\hat{\Omega}=0.1$	2.81	0	1.6×10^{-2}	0.76
	$\hat{\Omega}=0.03$	2.81	0	1.4×10^{-2}	0.76
GB+B	$\hat{\Omega}=1$	2.98	2.4×10^{-2}	5.3×10^{-2}	2.11
	$\hat{\Omega}=0.1$	2.95	4×10^{-3}	3.1×10^{-2}	2.29
	$\hat{\Omega}=0.03$	2.95	1.5×10^{-3}	2.6×10^{-2}	2.32
GB+QM+B	$\hat{\Omega}=1$	2.97	2.4×10^{-2}	5×10^{-2}	1.7
	$\hat{\Omega}=0.1$	2.95	4×10^{-3}	3×10^{-2}	1.8
	$\hat{\Omega}=0.03$	2.94	1.5×10^{-3}	2.4×10^{-2}	1.8

Table 3. Results for a "Comptonized" spectrum (eq. (6))

Set of data	"best" T (K)	"best" u	Δu (1σ)	$\chi^2_{\nu min}$
GB	2.73	0	5.6×10^{-2}	0.53
GB + QM	2.85	1.7×10^{-2}	1.3×10^{-2}	0.70
GB + B	2.93	2.4×10^{-2}	6.9×10^{-3}	1.84
GB + QM + B	2.92	2.0×10^{-2}	6.5×10^{-3}	1.55

and Perrenod (1977) who also showed that the allowed values of u are

consistent with a "Comptonization" of the CBR spectrum by a dense, very hot intergalactic gas which could account for the X-ray background in the range 2-100 keV. This could be indeed an exciting possibility but it must be noted that the large value of the associated $\chi^2_{\nu\min}$ suggests that the Berkeley group could have underestimated their errors. Of course, the larger the true errors are, the smaller is the statistical significance of the distortions.

If the Berkeley data are disregarded any significant evidence of distortions vanishes and one has, at the 3σ confidence level, $\mu_o \leq 1 \div 3 \times 10^{-2}$ (depending on $\hat{\Omega}$) and $u \leq 6 \times 10^{-2}$. The weak dependence of the upper limit of μ_o on $\hat{\Omega}$ can be somewhat surprising since the amplitude of the distortions increases as $\Omega^{-7/8}$ (eq. (3)). However an inspection of the data shows that there are very few (and not very accurate) measurements at $\lambda \gtrsim 3$ cm, where the maximum deviations from a BB spectrum are expected (eq. (4)). In fact no new measurements in RJ region have been carried out in the last ten years, and it is a pity since many valuable pieces of information could be hidden in this part of the CBR spectrum.

References

Avni, Y. 1976, Ap. J. 210, 642
Cash, W. 1976, Astron. Astrophys. 52, 307
Clegg, P.E. 1977, Lectures given at the Erice International School of Astrophysics
Danese, L., and De Zotti, G. 1977, Riv. Nuovo Cimento 7, 277
Danese, L., and De Zotti, G. 1978, Astron. Astrophys. 68, 157
Field, G.B., and Perrenod, S.C. 1977, Ap. J. 215, 717
Illarionov, A.F., and Sunyaev, R.A. 1974a, Astr. Zh. 51, 698 (Soviet Astr. 18, 413 (1975))
Illarionov, A.F., and Sunyaev, R.A. 1974b, Astr. Zh. 51, 1162 (Soviet Astr. 18, 691 (1975))
Robson, E.I., Vickers, D.G., Huizinga, J.S., Beckman, J.E., and Clegg, P.E. 1974, Nature 251, 591
Sunyaev, R.A., and Zeldovich, Ya.B. 1970, Ap. Space Sci. 7, 20
Woody, D.P., Mather, J.C., Nishioka, N.S., and Richards, P.L. 1975, Phys. Rev. Letters 34, 1036
Zeldovich, Ya.B., and Sunyaev, R.A. 1969, Ap. Space Sci. 4, 301

A UNIFIED TREATMENT OF DIFFERENT APPROACHES TO CLUSTERING OF GALAXIES

G. Dautcourt
Institute of Astrophysics, Academy of Sciences, Potsdam-Babelsberg

1. Introduction

The large-scale structure of distribution of galaxies following from optical observations has been studied by a fairly large number of different methods. The resulting picture turned out to be complicated. Apart from single clusters of galaxies, higher order clustering was suggested (Abell 1958, 1977, Kalinkov 1974).

Other methods led to the picture of a continuous clustering of galaxies with no preferred scales (Limber 1953, 1954, Layzer 1956, Peebles 1975, Peebles and Groth 1975, Peebles and Hauser 1974, Groth and Peebles 1977, Fry and Peebles 1978, Flin 1977, Rudnicki and Zięba 1978). Still open is the question if the observed picture of galaxy clustering reflects primordial structures (Zeldovich and Novikov 1975, Gott and Rees 1975, Zeldovich 1978) or arises from gravitational interactions independent of the initial state (Press and Schechter 1974, Doroshkevich and Zeldovich 1975, Fall and Saslaw 1976, Press and Lightman 1978, Silk and White 1978).

To compare the different approaches to galaxy clustering, a unified treatment would be very useful. It seems that the correlation function method widely used by Peebles and collaborators is able to provide a link between methods employed so far. There are other promising aspects of this type of approach: From models of formation of galaxies and clusters of galaxies one should be able to predict the parameters describing correlation functions of different order. Last not least dynamical problems can be treated by investigating the Liouville equation for the N-point probability distribution function in phase space, which is closely related to the correlation functions.

We shall not treat all these problems but concentrate on the relation between different measures of galaxy clustering. After introducing the basic notation in section 2, we discuss the frequency distribution of cell count, the joint distribution function for galaxy numbers in arbitrarily spaced cells, the "nearest neighbour" test, Zwicky's dispersion curve analysis, some aspects of the statistical

reduction method invented by the Kraków group, Turner and Gott's method to single out isolated galaxies and the usual serial correlation method. Section 4 discusses observations of galaxy clustering.

2. Statistical description of distribution of galaxies

2.1. Interacting point particles

The concepts used in recent years (for reviews of earlier results see Layzer 1959, de Vaucouleurs 1971) to describe a continuous clustering of galaxies are closely related to those introduced in the kinetic theory of interacting point particles (see, e.g., Montgomery and Tidman, 1964, or Klimontovich, 1964, 1975). One replaces the galaxies by point particles of equal mass m interacting through gravitational forces only. With an application to cosmology in mind one would like to have a general-relativistic treatment of gravitational interactions. However, most existing formalisms are based on a Hamiltonian description with the interactions entering explicitly in a action-at-a-distance manner. It is difficult to carry over these concepts to general relativity, where interactions are primarily given implicitly in terms of field laws. So far only the Klimontovich formalism was extended for the curved space-time of general relativity. Here we follow an approximate procedure starting from a purely Newtonian description and adding later the effects of expansion by introducing comoving coordinates (Saslaw 1972, Yahil 1976, Fall and Saslaw 1976, Inagaki 1976, Davis and Peebles 1977). This should be a sufficient approximation provided characteristic distances like the correlation lengths are all small compared with the Hubble distance.

Consider a gas of N point particles of equal masses m enclosed in a large box of volume V. Following Gilbert (1971), we introduce the Liouville probability density $f(1,2,\ldots N)$ (denoted as f_N in Dautcourt 1977), where $f(1,2,\ldots N)d(2)\ldots d(N)$ is the probability to find particle 1 in a $d\underline{r}_1$ - neighbourhood of a point with the cartesian coordinates $\underline{r}_1$ and within $d\underline{v}_1$ of velocity $\underline{v}_1$, particle 2 in a $d\underline{r}_2$-neighbourhood of $\underline{r}_2$ and around $\underline{v}_2$ etc. For simplicity we have abbreviated $d\underline{r}_i d\underline{v}_i$ by d(i) (we also use dx_i and x_i for d(i) and $(\underline{r}_i,\underline{v}_i)$ respectively). $f(1,2,\ldots N)$ is assumed to be symmetric in all arguments and to give 1 if integrated over the whole phase space:

$$f(1,2,\ldots i,\ldots k,\ldots N) = f(1,2,\ldots k,\ldots i,\ldots N) \qquad (2.1)$$

$$\int f(1,2,\ldots N)\, d(1)\ldots d(N) = 1 \qquad (2.2)$$

The probability distribution $f_k(1,\ldots k)$ of coordinates and velocities of k particles irrespective of the other particles (that is, averaged over all possible states of the N-k remaining particles) is correspondingly given by

$$f_k(1,2,\ldots k)/V^k = \int f(1,2,\ldots N)\, d(k+1)\ldots d(N) \tag{2.3}$$

with k running from 1 to N - 1. (The volume factor is introduced for normalization).

The meaning of the Liouville probability density becomes clear if a whole ensemble of boxes of equal volumes V but different configurations of particles is considered. The quantities of most interest are ensemble averages. For instance, the phase space particle density, defined as a sum of Dirac delta functions:

$$n(x) = \sum_{i=1}^{N} \delta(x - x_i) \tag{2.4}$$

has the ensemble average

$$\langle n(x) \rangle = \sum_{i=1}^{N} \delta(x-x_i)\, f(1,2,\ldots N)\, d(1)\ldots d(N) = n_o\, f(1) \tag{2.5}$$

with $n_o = N/V$ as mean particle density and $f_1(1) \equiv f(1)$ as single particle function. The ensemble average of the product of particle densities $n(x_1)$, $n(x_2)$ and $n(x_3)$ at different points x_1, x_2, x_3 in phase space is

$$\begin{aligned}
&\langle n(x_1)\, n(x_2) \rangle = n_o^2\, f_2(1,2) + n_o\, \delta(x_1-x_2)\, f(1) \\
&\langle n(x_1)\, n(x_2)\, n(x_3) \rangle = n_o^3\, f_3(1,2,3) + n_o^2 \left[\delta(x_1-x_2)\, f_2(2,3) \right. \\
&\quad \left. + \delta(x_1-x_3)\, f_2(1,2) + \delta(x_2-x_3)\, f_2(1,3) \right] \\
&\quad + n_o\, \delta(x_1-x_2)\, \delta(x_2-x_3)\, f(1)
\end{aligned} \tag{2.6}$$

We may also correlate the density fluctuations $\delta n(x_1) = n(x_1) - \langle n(x_1) \rangle$ at different phase space points:

$$\begin{aligned}
&\langle \delta n(x_1) \rangle = 0, \\
&\langle \delta n(x_1)\, \delta n(x_2) \rangle = n_o^2\, g_2(1,2) + n_o\, \delta(x_1-x_2)\, f(1), \\
&\langle \delta n(x_1)\, \delta n(x_2)\, \delta n(x_3) \rangle = n_o^3\, g_3(1,2,3) + n_o^2 \times \\
&\times \left[\delta(x_1-x_2)\, g_2(2,3) + \delta(x_1-x_3)\, g_2(1,2) + \delta(x_2-x_3)\, g_2(1,3) \right] \\
&\quad + n_o\, \delta(x_1-x_2)\, \delta(x_2-x_3)\, f(1).
\end{aligned} \tag{2.7}$$

The functions $g_i(1,2,\ldots i)$ introduced here are called correlation functions. They may be defined also by

$$f_2(1,2) = f(1)\, f(2) + g_2(1,2),$$

$$f_3(1,2,3) = f(1)\, f(2)\, f(3) + f(1)\, g_2(2,3) + f(2)\, g_2(1,3) + f(3)\, g_2(2,2) + g_3(1,2,3), \tag{2.8}$$

etc.
Note that the phase space integral over any argument of a correlation function vanishes:

$$\int g_i(1,\ldots k,\ldots i)\, d(k) = 0. \tag{2.9}$$

The probability distribution function $f(1,\ldots N)$ satisfies a Liouville equation

$$\frac{\partial f}{\partial t} + \sum_{i=1}^{N} \underline{v}_i \frac{\partial f}{\partial \underline{x}_i} + \sum_{i,j=1}^{N} \underline{b}_{i,j} \frac{\partial f}{\partial \underline{v}_j} = 0. \tag{2.10}$$

From (2.10), equations for the time derivatives of correlation functions may be derived. We stop at this point and turn to the problem of spatial distribution of galaxies.

2.2. The space and surface distribution of galaxies

If dynamical questions are not considered, we may integrate away all velocity dependence in $f_i(1,\ldots i)$ as well as in $f(1\ldots N)$. In all of the following formulas it is assumed that this integration has been carried out. x_1 or dx_1 (or also $d(1)$) is now taken as abbreviation for $\underline{x}_1$ or $d\underline{x}_1$. Then (2.6) - (2.9) hold also for the velocity integrated terms.

A useful concept is the Mayer cluster expansion of the Liouville distribution function f. If the galaxies are randomly distributed with no correlation present, f is separable into a product of 1-point functions $V^{-N} \prod_{i=1}^{N} f(i)$ (apart from a normalizing factor). When correlations are taken into account, additional terms appear:

$$f(1,\ldots N) = V^{-N} \left\{ \prod_{i=1}^{N} f(i) + \sum_{i\neq 1} g(i,j) \prod_{k=1}^{N}{}^{(i,j)} f(k) + \sum_{i\neq j\neq k} g_3(i,j,k) \prod_{l=1}^{N}{}^{(i,j,k)} f(l) + \ldots \right.$$

$$+ \sum_{i_1 \neq i_2 \cdots \neq i_{N-1}} g_N(1,\ldots i_{N-1})\ f(N) \Big\} \qquad (2.11)$$

The equation (2.8) may be considered as particular case of the representation (2.11) (Note that f_1, g_1 are defined as dimensionless quantities, while f carries a dimension).

Thus far we have described the space distribution of galaxies. The projected two-dimensional galaxy distribution on, e.g., a photographic plate can be treated quite similarly. Using as a rule capital letters for surface quantities, we write

$$F(1,\ldots N)\ dX_1 \ldots dX_N$$

for the probability to find galaxy 1 in a two-dimensional dX_1-cell around X_1, galaxy 2 in a dX_2-cell around X_2 etc. All equations (2.1)-(2.9) may be written as relations for surface quantities by changing to capital letters and replacing V by 4π, the surface of the unit sphere. We list the most frequently employed equations:

$$\begin{aligned}
&\langle \delta \mathcal{N}(X_1) \rangle = 0 \\
&\langle \delta \mathcal{N}(X_1)\ \delta \mathcal{N}(X_2) \rangle = \mathcal{N}^2 G_2(1,2) + \mathcal{N} \delta(X_1 - X_2)\ F_1(1) \\
&\langle \delta \mathcal{N}(X_1)\ \delta \mathcal{N}(X_2)\ \delta \mathcal{N}(X_3) \rangle = \mathcal{N}^3 G_3(1,2,3) \\
&+ \mathcal{N}^2 \left[\delta(X_1 - X_2)\ G_2(1,3) + \delta(X_1 - X_3)\ G_2(1,2) + \delta(X_2 - X_3)\ G_2(1,2) \right] \\
&+ \mathcal{N} \delta(X_1 - X_2)\ \delta(X_2 - X_3)\ F_1(1).
\end{aligned} \qquad (2.12)$$

The Mayer cluster expansion for the surface is

$$F(s,\ldots N) = \frac{1}{(4\pi)^N} \left\{ \prod_{i=1}^{N} F_1(1) + \sum_{i=j} G_2(i,k) \prod_{i=j}^{N} (i,j)\ F_1(k) + \ldots \right\} \qquad (2.13)$$

In (2.12), (2.13) N is the mean number of galaxies per steradian corresponding to a given photographic plate. The relation between the space and surface correlation functions is discussed in section 4.1.

3. Classical measures of galaxy clustering

To determine the correlation functions of all higher orders directly from observations is not very practical and requires a lot of computer time. However, many other measures of the clustering of galaxies have been proposed, which are sensitive to higher-order corre-

lations and often easier to calculate. In every case one should be able to find the connexion with the correlation functions.

3.1. Frequency distribution of cell counts

An often employed measure of galaxy clustering is the observed frequency distribution of numbers of galaxies in equal cells on the photographic plate (Hubble 1934, Bok 1943, Abell 1958, Zonn 1968, Flin et al. 1974, Dodd et al. 1975). If galaxies are clustered one expects that the observed number $\nu(i)$ of cells with a very small or very large population i is higher than the number corresponding to a random (that is, Poisson) distribution of galaxies. Let $\langle N \rangle$ be the mean number of galaxies per cell, N_{tot} the total number of galaxies in all cells, then

$$\nu_o(i) = N_{tot} \frac{\langle N \rangle^i e^{-\langle N \rangle}}{i!} \tag{3.1}$$

would correspond to a random distribution. The deviation of the actual distribution from a random one could be measured by the quantity

$$\chi^2 = \sum_{i=0}^{N_{max}} \frac{(\nu[i] - \nu_o[i])^2}{\nu_o[i]} \tag{3.2}$$

(N_{max} is the largest number of galaxies in the cells of a given sample). The probability that a random sample drawn from a Poisson distribution for a given value of χ^2 equal to or exceeding that calculated from (3.2) is the integral over the χ^2 distribution function (Abell 1958).

The observed frequency distribution $\nu(i)$ contains more information. We may express the ensemble average of $\nu(i)$ in terms of cell moments of the correlation function. The calculations leading to this result are typical for the treatment of galaxy clustering given here, thus we give a short sketch. We want to calculate the probability P(i) that in a square cell $l \times l = \omega$ the number of counted galaxies is just i. The probability that the i galaxies numbered by 1 through i out of the total number N (N is the number of galaxies which could be seen over the whole sky up to the limiting magnitude of the given plate) are found in the cell, while the remaining galaxies with numbers i+1 through N are outside the cell is given by

$$p_1 = \int F(1,\ldots i,i+1,\ldots N)\, dX_1 \ldots dX_i dX_{i+1} \ldots dX_N .$$

Here the integration with respect to X_1 through X_i extends over the cell, whereas the integration with respect to the remaining variables extends over the whole sky outside the cell. Since any other collection of i galaxies out of the total population of N galaxies has the same probability p_1 to occur, and since there are $\binom{N}{i}$ ways to choose such a collection, the probability P(i) as defined above is given by

$$P(i) = \binom{N}{i} p_1 .$$

For a random distribution of galaxies, F is the product of 1-point distribution functions f_1, apart from a factor $1/(4\pi)^N$. Taking $f_1=1$ (which corresponds to a homogeneous galaxy distribution on the plate), one obtains, carrying out the integrations and writing P_o instead of P:

$$P_o(i) = \binom{N}{i} \omega^i (4\pi - \omega)^{N-i} (4\pi)^{-N} .$$

Since $\omega/4\pi = \bar{N}/N$, where $\bar{N}$ is the mean number of galaxies in a cell, we may also write, using Stirling's formula

$$i! = \sqrt{2\pi N}\, N^N e^{-N}$$

for large N, and letting $N \longrightarrow \infty$,

$$P_o(i) = \frac{\langle N \rangle^i}{i!} e^{-\langle N \rangle} . \qquad (3.5)$$

This is the Poisson distribution, as expected. There are correction terms if the cell is so large that $i \ll N$ is no longer valid. Usually, these corrections can be neglected. More important are corrections arising from the presence of correlations. If the full Mayer cluster expansion is employed instead of its first term, a similar type of calculation leads to the expression

$$P_1(i) = P_o(i)\left(1 + \sum_{r=2}^{N} \mathcal{T}_r\right) \qquad (3.6)$$

with

$$\mathcal{T}_r = i!\, G_{ro} \sum_{s=o}^{r} \frac{(-1)^s \langle N \rangle^s}{(i-r+s)!\,(r-s)!\,s!} \qquad (3.7)$$

G_{ro} is the r-point correlation function integrated over a cell (Appendix). For weak clustering the maximum of P(i) compared with that of $P_o(i)$ is shifted to smaller values of i. Also the expected excess of galaxies with small and large i follows from the formula (3.6). In Fig. 1 the data from the Jagiellonian field (Rudnicki et al. 1973) are

fitted to a relation (3.6) with only π_2 taken into account.

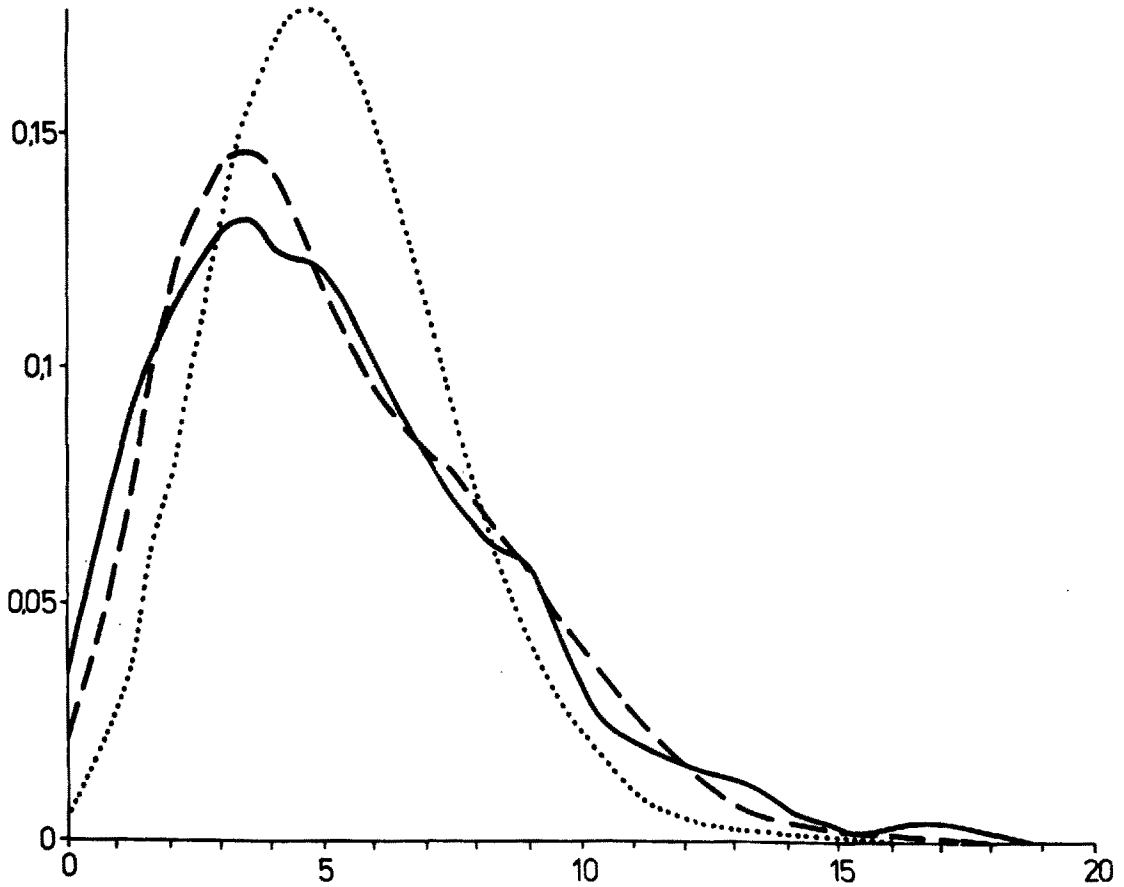

Fig. 1. Frequency distribution of counting cell (7!5 × 7!5) population for the Jagiellonian field, together with a Poisson distribution (dots) corresponding to the same mean galaxy number 5.2 per cell, and a distribution with only the second-order moment $G_{2o} = 0.132$ (obtained from the observational data by fitting equation (3.6) with the constraint $G_{ro} = 0$ for $r > 2$) taken into account (dashes)

A better fit of the observational data is obtained if higher-order correlation moments are considered (Fig. 2). If the clustering is strong, that is, if the higher order moments G_{ro} with $r > 2$ are of the same order as G_{2o}, the picture changes.

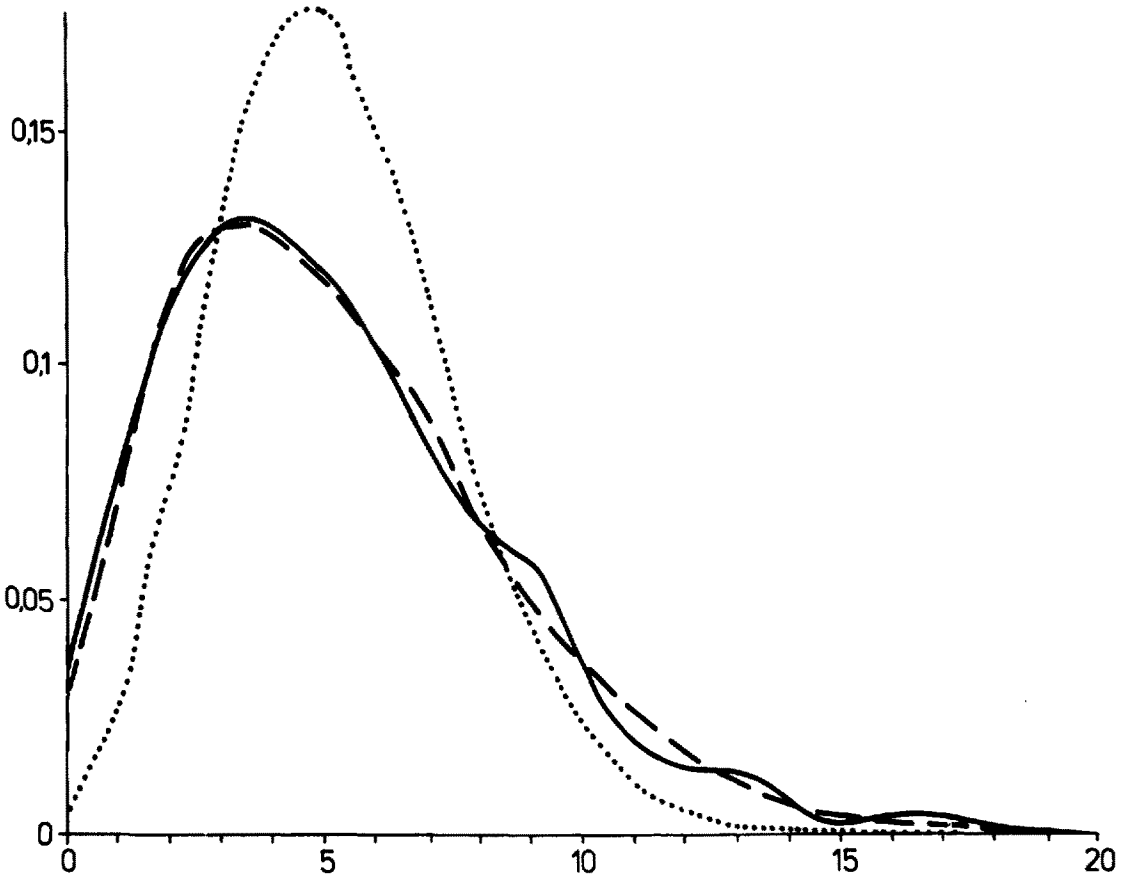

Fig. 2. Same as Fig. 1, but with cell moments up to the order 4 in the dashed curve (G_{2o} = 0.174, G_{3o} = 0.019, G_{4o} = 0.046)

In particular, several maxima are predicted to be present (Fig. 3). We do not recommend to use this method to determine the cell moments, since the moments can be found much easier directly from the observational data. However, the plot of the frequency distribution gives an immediate impression, what type of clustering (weak or strong) might be present.

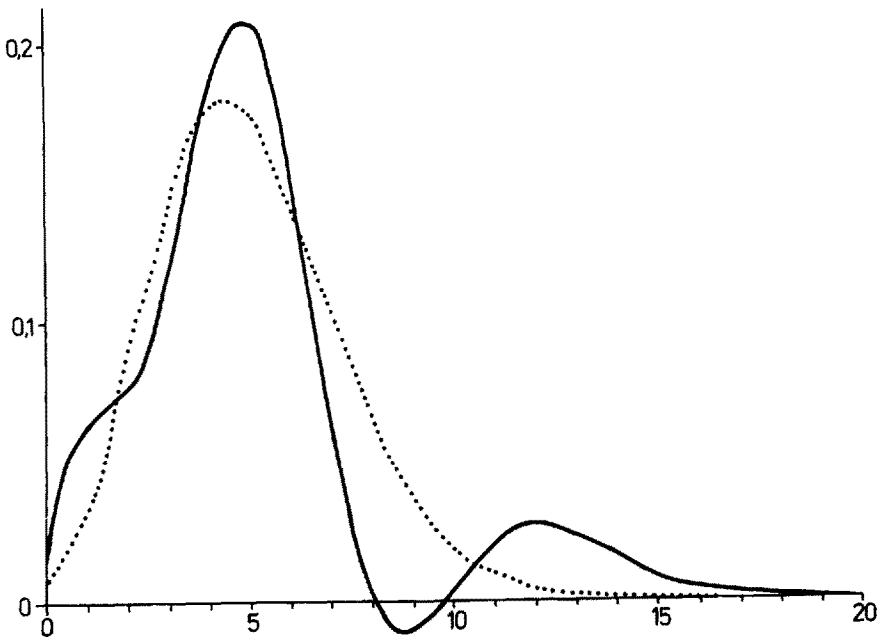

Fig. 3. Poisson frequency distribution of counting cell population for a mean number galaxy per cell corresponding to the Jagiellonian field (dotted) together with a distribution obtained from (3.6) with $G_{2o} = G_{3o} = G_{4o} = 0.2$ (remaining G_{ro} taken zero)

3.2. The joint distribution function for numbers of galaxies in arbitrarily spaced cells

A further useful quantity is the joint distribution function of the cell population, defined as the unconditioned probability

$$P = P_k(i_1, i_2 \cdots i_k),$$

that any first cell of arbitrary shape contains just i_1 galaxies, a second cell i_2 galaxies etc, the k-th cell i_k galaxies. The special case $P_1(i)$ is identical with the distribution function of the cell population discussed in the last section. If no clustering is present, P_k is given by the product

$$P_k(o) = \prod_{l=1}^{k} \frac{\langle N_l \rangle^{i_l}}{i_l!} e^{-\langle N_l \rangle} \tag{3.8}$$

of k Poisson distributions. Again, clustering increases in general the probability for small as well as for large i. A knowledge of P_k allows to answer many questions connected with the surface distribution of galaxies. We note few of them: An inspection of photographic plates often shows small blank areas where no galaxies are seen up to the limiting magnitude of the plate. A clustered distribution of galaxies produces blank regions with higher probability than a distribution by chance. But intergalactic absorption clouds may also be present. To decide between these possibilities, it would be important to know the probability that a certain grouping of blank areas (or perhaps weakly populated areas) arises from the existence of clustering (for a single blank area equation (3.6) applies). Another example: Large-scale clustering may produce variations in the mean density of galaxies in several parts of a photographic plate. But local variations of the plate sensitivity may produce similar effects. Again it is interesting to know the probability that a given excess of galaxy numbers may occur already by chance or as a consequence of galaxy clustering.

The required calculation is a straightforward generalization of that given in the last section. We present the results only for the case k = 2 and for correlation functions up to the third order:

$$\begin{aligned}
P_2(i,k) = P_2^o(i,k) \Big\{ & 1 + \tfrac{1}{2} G_{2o}^{(1)} \left[\bar{i}^2 - 2i\bar{i} + i(i-1) \right] \\
& + \tfrac{1}{2} G_{2o}^{(2)} \left[\bar{k}^2 - 2k\bar{k} + k(k-1) \right] + G_{2o}^{(1,2)} \left[\bar{i}\,\bar{k} + ik - i\bar{k} - k\,\bar{i} \right] \\
& + \tfrac{1}{6} G_{3o}^{(1)} \left[-\bar{i}^3 + 3i\bar{i}^2 + i(i-1)(i-2) - 3i(i-1)\bar{i} \right] \\
& + \tfrac{1}{6} G_{3o}^{(2)} \left[-\bar{k}^3 + 3k\bar{k}^2 + k(k-1)(k-2) - 3k(k-1)\bar{k} \right] \\
& + \tfrac{1}{2} \int_{\omega_2} G_{31}^{(1)}(X_1) \frac{dX_1}{\omega_2} \left[-\bar{i}^2\bar{k} + 2i\bar{i}\bar{k} + k\bar{i}^2 - 2ik\bar{i} - i(i-1)\bar{k} \right. \\
& \left. + ik(i-1) \right] + \tfrac{1}{2} \int_{\omega_1} G_{31}^{(2)}(X_1) \frac{dX_1}{\omega_1} \left[-\bar{k}^2\bar{i} + i\bar{k}^2 + 2\bar{i}\bar{k}k - 2ik\bar{k} \right. \\
& \left. - k(k-1)\bar{i} + ik(k-1) \right] \Big\}
\end{aligned} \tag{3.9}$$

We have written $\bar{I} = \mathcal{N}\omega_1$, $\bar{k} = \mathcal{N}\omega_2$ for the mean number of galaxies in the cells 1 and 2. Furthermore, we distinguish between the moments for the two cells (which are different since $\omega_1 \neq \omega_2$ in general). Another notation is

$$G_{20}^{(1,2)} = \int_{\omega_1} \int_{\omega_2} G_2 (X_1, X_2) \frac{dX_1}{\omega_1} \frac{dX_2}{\omega_2} .$$

If the angular distance Θ between the cell centers is large compared with the cell size, then $G_{20}^{(1,2)} \simeq \omega(\Theta)$ is a good approximation.

Evidently, if one sums (3.9) over all k one should obtain $P_1(i)$. This may be checked by direct calculation.

3.3. Nearest neighbor test

The distribution of the distances to the first, second etc. neighbor of a galaxy may be used as another measure of galaxy clustering. Provided galaxy 1 is at the position X_1, then the probability distribution function for all other galaxies is given by $4\pi F(1,\ldots N)/F_1$ instead of F. The probability that the next galaxy is at an angular distance between Θ and $\Theta + d\Theta$ is given by:

$$D(\Theta)\, d\Theta = 4\pi (N-1)\, d\Theta \sin\Theta \int d\varphi \int F(1,\ldots N) dX_3 \ldots dX_N , \qquad (3.10)$$

where the first integration is over the circumference of the Θ-circle, and the other integrations are over the whole sky outside the Θ-circle. Working out the integrals with the Mayer cluster expansion for F one obtains

$$D(\Theta) = 2\pi \sin\Theta\, \mathcal{N}\, e^{-\langle N\rangle} \left[1 + \sum_{r=2}^{N} \vartheta_r\right] \qquad (3.11)$$

with

$$\vartheta_r = \frac{\langle N\rangle^r (-1)^r}{r!} G_{ro} \langle N\rangle^{r-1} \frac{(-1)^{r-1}}{(r-1)!} G_{r1}(X_1) \qquad (3.12)$$

$$+ \frac{\langle N\rangle^{r-1}(-1)^{r-1}}{(r-1)!} \int \frac{d\varphi}{2\pi} G_{r1}(X_2) + \frac{\langle N\rangle^{r-2}(-1)^{r-2}}{(r-2)!} \int \frac{d\varphi}{2\pi} G_{r2}(X_1, X_2)$$

where the integrals must be taken over the circumference of the Θ-circle.

In a similar way one may calculate the distribution of the distances to the second, third etc neighbor of the test galaxy. If the mean number $\langle N\rangle$ of galaxies in a Θ-circle is small compared with 1, then approximately $D(\Theta) \simeq D_0(\Theta)(1 + w(\Theta))$, where D_0 is the correspond-

ing distribution function for a randomly distributed sample of galaxies (cf. (3.11)). Thus, for $\langle N \rangle \ll 1$, $D(\theta)$ is larger than D_0. For $\langle N \rangle \gtrsim 1$ the situation is more complicated.

In connexion with the distribution of galaxies, empirical estimates of the function $D(\theta)$ were discussed by Wagoner (1967), Wagoner and Bogart (1973), Peebles and Soneira (1977) and by Kalinkov (1978).

3.4. Cell moments of second order

The dispersion of numbers of galaxies in square cells, $\sigma^2 = \langle (N - \langle N \rangle)^2 \rangle$ (where $N = \int \mathcal{N} dX$ is integrated over a cell and $\langle N \rangle = \mathcal{N}\omega$ the mean number of galaxies in the cell of area ω) taken as a function of cell size $l = \omega^{1/2}$ gives a measure of galaxy clustering which is easily estimated from the observational data. With equation (2.12) one obtains

$$\sigma^2 = \langle (N - \langle N \rangle)^2 \rangle = \langle N \rangle (1 - \varepsilon_1) + \langle N \rangle^2 G_{20}(l) \quad (3.13)$$

We have included here a factor $(1 - \varepsilon_1)$ for the following reason: The surface two-point correlation function $G_2(1,2)$ contains a Dirac delta function term $\sim \delta(X_1 - X_2)$ resulting from the spread in the luminosity function of galaxies (section 4.1). The second order cell moment $G_{20}(l)$ may be written

$$G_{20}(l) = \int \frac{w(\theta)}{\omega^2} dX_1 dX_2 = k_0 \, w(l) \quad (3.14)$$

(note $w(\theta)$ does not contain the delta function term by definition). In the case of a power law $w \sim \theta^{-\delta}$ for $\theta \leq l$, the factor k_0 is independent of the cell size l but depends on δ (Table 1).

Table 1. Smoothing factors k for a square cell (equation (3.14)), for different values of the index δ in $w(\theta) \sim \theta^{-\delta}$, obtained by Monte Carlo integration

δ	0.5	0.6	0.7	0.8	0.9	1.0	1.2
k_0	1.583	1.764	1.980	2.240	2.557	2.950	4.073

Usually the root $\varkappa$ of the ratio $\varkappa^2$ of the observed dispersion σ^2 to that expected for a Poisson distribution, $\sigma_0^2 = \langle N \rangle$, is plotted as a function of cell size l (see, e.g., Zwicky 1957). Thus, for a power law $w \sim \theta^{-\delta}$

$$\varkappa^2 = 1 - \varepsilon_1 + k\, w_0\, \mathcal{N}\, l^{2-\delta} \qquad (3.15)$$

Provided $\delta < 2$, $\varkappa$ is an increasing function of cell size. This is also observed in most cases (see, e.g., Fig. 11 in Dautcourt 1977, referring to data from the Jagellonian catalog). Thus the dispersion curve $\varkappa(l)$ considered as function of cell size gives essentially the some information as the correlation function $w(\vartheta)$.

3.5. Parameters connected with the statistical reduction method

A statistical method to investigate the degree and type of galaxy clustering was developed by the Kraków group (Rudnicki et al. 1973, Flin et al. 1974, Flin 1974, Zieba 1975, Flin 1977, Rudnicki and Zieba 1978). The reduction method shares the advantage of other measures of galaxy clustering to condense the structural properties of the galaxy distribution into few parameters, here called the index of concentration c, of grouping g, and of weak and strong anisotropy. The latter parameters describe a possible anisotropy (directional dependence) in the distribution of galaxies. They are not accessible within our framework since we assume statistical homogeneity of the surface random process.

The concentration index c indicates a deficiency or an excess of highly populated cells compared with the numbers of cells for a random distribution. In our notation we may write (cf. Flin 1974):

$$c = \frac{1}{\langle N\rangle^2 (1 - 1/\eta \langle N\rangle)} \sum_{i=2}^{\infty} i(i-1) P_1(i), \qquad (3.16)$$

where η is the total number of elementary cells and $\langle N\rangle$ the mean number of galaxies in a cell. The grouping index may be calculated from

$$g = \frac{1 - \langle N\rangle/\eta}{1 - c\langle N\rangle/\eta} \frac{1}{\langle N\rangle^2 (1 - 1/\eta \langle N\rangle)} \sum_{i,k}^{\infty} ik\, P_2(i,k) \qquad (3.17)$$

g measures - if $g > 1$ - the tendency for highly populated areas to gather together (or to avoid each other, if $g < 1$).

With the formula derived in 3.1 and 3.2 for P_1 and P_2, one obtains from the definition (3.16) and (3.17) up to (and including) correlations of order r = 4, if $1/\eta \langle N\rangle$ is neglected compared with 1, the simple result

$$c = 1 + G_{20}, \qquad (3.18)$$

the excess of the concentration index over 1 is equal to the second order cell moment. It is likely that the cancellation of terms from higher-order correlations continues for all orders $r > 4$, but a proof remains to be given. Similarly one obtains from (2.13) for the grouping index, again neglecting terms $\langle N \rangle / \eta \quad c \langle N \rangle / \eta$, $1/(\langle N \rangle \eta)$ compared to 1, after some calculation

$$g = 1 + \int G_2(1,2) \frac{dX_1 dX_2}{\omega_1 \omega_2} \tag{3.19}$$

Terms resulting from third order correlations do not occur in (3.19). Again one suspects that this also holds for higher-order correlations. If the cells with areas ω_1 and ω_2 are separated an angular distance θ large compared with the cell size, one has with a good accuracy

$$g \simeq 1 + w(\theta) \tag{3.20}$$

Thus - at least in the case of weak clustering - the concentration index as well as the grouping index measure well-known quantities. This agrees with the qualitative descriptions given for these indices. A first glance on the data from the Jagellonian catalog shows also numerically order-of-magnitude agreement. However, a detailed comparison between the data obtained from the statistical reduction method and those of the correlation function approach has still to be carried out.

3.6. Single galaxies

Turner and Gott (1975) have pointed out that the distribution of galaxies brighter than 14th magnitude in the Zwicky cataloge (Zwicky et al. 1961-68) shows a remarkable feature. If galaxies with no neighbor within a circle of radius θ_1 (= 45′) are correlated with all galaxies, the corresponding correlation function $w_s(\theta)$ for "singles" is nearly zero for scales $\theta > \theta_1$ up to $\theta = 15^\circ$. Thus the population of "single" galaxies appears to be a well-defined sample of uniformly distributed galaxies. It was later suggested by Fall et al. (1976) that the flatness of the two-point correlation function for singles is probably an artifact of the selection criteria used to define a "single" galaxy. This means that the existence of isolated galaxies may not contradict the picture of continuous clustering.

Our approach is well suited to handle problems of this type. With Turner and Gott we define the "singles" correlation function

$w_s(\Theta)$ by

$$w_s(\Theta) = \frac{P_s(\Theta)}{P_o(\Theta)} - 1, \tag{3.21}$$

where $P_s(\Theta)$ is the probability to find a galaxy of any type (supposed to be at a sky position X_2) at an angular distance Θ from a single galaxy (at sky position X_1) and $P_o(\Theta)$ is the expected probability if the galaxies are randomly distributed. $P_s(\Theta)$ is easily calculated from (2.13) as

$$P_s(\Theta) = P_o(\Theta) \left\{ 1 + \sum_{r=2}^{N} \frac{(-1)^r N_1^{r-2}}{r!} \left[r(r-1) G_{r2}(X_1, X_2) - r N_1 G_{r1}(X_1) \right. \right.$$

$$\left. \left. - r N_1 G_{r1}(X_2) + N_1^2 G_{ro} \right] \right\}, \tag{3.22}$$

where $N_1 = \pi \Theta_1^2 \mathcal{N}$ is the mean number of galaxies in a Θ_1-circle. Taking only the two-point correlation function into account, one obtains, if the moment G_{20} over a Θ_1-circle is replaced by $\tilde{k}_o w(\Theta_1)$,

$$w_s(\Theta) = w(\Theta_1) \left[N_1^2 \tilde{k}_o / 2 - 2 N_1 / (2 - \delta) \right] + w(\Theta)(1 - N_1) \tag{3.23}$$

Clearly, equation (3.23) is not sufficient to explain the data from the Turner-Gott sample of galaxies at spacings near $\Theta \simeq \Theta_1$, since higher-order correlations are not taken into account. But even with (3.23), the decrease of the "single-to-total" galaxy correlation function $w_s(\Theta)$ with increasing Θ_1 (hence increasing N_1) found by Soneira and Peebles (1977) could be understood qualitatively. $N_1 = 1$ or $\Theta_1 = 79'$ would make $w_s(\Theta)$ flat for all Θ, leaving only the constant first term. The first term in (3.23) is always negative, which clearly shows that higher-order correlations must be considered (the expressions resulting from 3-point correlations also contain a - positive, constant term).

3.7. Direct estimates of correlation functions

Most methods characterizing clustering of galaxies considered in the previous sections involve correlation functions of all orders. A clean separation between different orders can easily be obtained. The probability to find a galaxy in a cell dX_2 around X_2, provided there is already a galaxy at X_1, is given by

$$P\, dX_2 = 4\pi dX_2 \int F_N(X_1, X_2 \ldots X_N) / F_1(X_1) \cdot dX_3 \ldots dX_N , \tag{3.24}$$

where the integration is to be carried out over the whole surface. Introducing the cluster expansion, it is immediately seen that all terms with correlation functions G_i, $i \geqslant 3$ drop out, because the integration of a correlation function over the whole sky gives zero. Thus we are left with

$$P\, dX_2 = \left[1 + G_2(X_1, X_2) \right] dX_2 \,. \tag{3.25}$$

In a similar way we obtain correlation functions of higher order. For example, the probability to find a galaxy in a cell dX_2 around X_2 and a galaxy in a dX_3 cell around X_3, provided there is a galaxy at X_1 can be worked out as

$$\begin{aligned} P\, dX_2 dX_3 = \big[& 1 + G_2(X_1, X_2) + G_2(X_1, X_3) + G_2(X_2, X_3) \\ & + G_3(X_1, X_2, X_3) \big] \, dX_2 dX_3 \,, \end{aligned} \tag{3.26}$$

since again higher-order terms vanish. Estimates of the probabilities from empirical data allow to determine the correlation functions. The most frequently employed method however, uses galaxy counts $N = \int \mathcal{N} dX$ in many cell pairs, separated by a fixed angular distance Θ_{12}, to determine two-point correlations, or counts in cell tripels with fixed angular distances Θ_{12}, Θ_{13}, Θ_{23} for three-point correlations. This allows to estimate the left hand side of

$$\langle (N_1 - \langle N \rangle)(N_2 - \langle N \rangle) \rangle = \langle N \rangle^2 \int w(\Theta) \frac{dX_1 dX_2}{\omega_2} \simeq \langle N \rangle^2 \bar{w}(\Theta) \tag{3.27}$$

directly from observations. Doing this, we implicitly assume that the surface random process is homogeneous and isotropic (see the Appendix). Since one obtains from (3.27) only correlation functions smoothed over a cell area, a problem arises if the cell distances are comparable with the cell size ℓ. A deconvolution would require to know the counts in cells with sizes smaller than ℓ. However, the effects are small and need only to be taken into account for adjacent cells. Let us define a smoothing correction factor $k_\Theta(\ell)$ by

$$w(\Theta) = \bar{w}(\Theta) / k_\Theta(\ell) \,.$$

Then an easy numerical integration gives the values in Table 2. Note that $k_\Theta(\ell)$ depends only on δ and on the ratio Θ/ℓ, if the same power law applies at all scales $\Theta \leqslant \ell$ (this was assumed for the calculation of Table 2).

Table 2. Smoothing correction factors $k_\vartheta(l)$ for square cells $l \times l$ for two-point correlations $\sim \vartheta^{\delta}$ as a function of cell distance ϑ/l (cell distance in units of cell size) and δ

ϑ/l \ δ	0,5	0,6	0,7	0,8	0,9	1,0	1,2
1	1,022	1,034	1,049	1,066	1,087	1,113	1,178
2	1,005	1,007	1,010	1,014	1,017	1,021	1,031
3	1,002	1,003	1,005	1,006	1,008	1,009	1,014
4	1,001	1,002	1,003	1,003	1,004	1,005	1,008
5	1,001	1,001	1,001	1,002	1,003	1,003	1,005

4. The space distribution of galaxies derived from various catalogs

4.1. Connexion between the space and surface density of galaxies

The well-known formula for the number of galaxies, per steradian, brighter than a given apparent magnitude m,

$$\mathcal{N}_c(m) = \int_0^r \frac{r^2 dr\, n(r)\, \vartheta_c(M\,[r])}{\sigma^3 (1+z)^3}, \qquad \sigma \equiv (1 + \tfrac{kr^2}{4}), \tag{4.1}$$

where $n(r) = n_0(1+z)^3$ is the space density of galaxies, holds also for the corresponding fluctuations δn and $\delta\mathcal{N}_c$. Note

$$\vartheta_c(M) = \int_{-\infty}^{M} \vartheta(M)\,dM \tag{4.2}$$

in (4.1) is the fraction of galaxies with absolute magnitudes brighter that M. Thus from (4.1) the average $\langle \delta\mathcal{N}_1\ \delta\mathcal{N}_2 \rangle$ of the product of fluctuations at different sky directions 1 and 2 can be expressed in terms of the space correlation function $\langle \delta n_1\ \delta n_2 \rangle$:

$$\langle \delta\mathcal{N}_1\ \delta\mathcal{N}_2 \rangle = \int_0^\infty \int_0^\infty \frac{dr_1 dr_2\ r_1^2 r_2^2\ \vartheta_c(M_1)\vartheta_c(M_2)}{\sigma_1^3\ \sigma_2^3\ (1+z_1)^3 (1+z_2)^3} \langle \delta n_1\ \delta n_2 \rangle \tag{4.3}$$

(We have assumed as usual that the luminosity function is universal and that the luminosities are not correlated with the position of galaxies). To calculate an expression like (4.3), we use the property $\langle \delta n_1\ \delta n_2 \rangle$ to decrease rapidly for large values of the separation between two space points 1 and 2. Furthermore, one has to take into account that the coordinates x in (2.4) - (2.10) are local (not co-

moving) coordinates, whereas r in (4.1) is the comoving radial coordinate in the Friedman line element

$$ds^2 = -dt^2 + \frac{R^2(t)}{\mathcal{G}^2}\,(dr^2 + r^2\, d\Omega^2) \tag{4.4}$$

Changing the integration coordinates one may rewrite (4.3) in the form of the second equation in (2.12):

$$\langle \delta\mathcal{N}(X_1)\ \ \delta\mathcal{N}(X_2)\rangle \;=\; \mathcal{N}^2 G_2(1,2) \;+\; \mathcal{N}\delta(X_1-X_2)\ F_1(1) \tag{4.5}$$

with

$$G_2(1,2) = \frac{2\ n_o^2}{\mathcal{N}^2}\ \int_0^\infty \frac{dr(1+z)r^4}{\mathcal{G}^5}\ \theta_c^{\,2}(M)\ \gamma_2\left(\frac{c\theta\phi}{H_o}\right) - \frac{\varepsilon_1\ \delta(X_1-X_2)}{\mathcal{N}} \tag{4.6}$$

as the generalized Limber equation (Limber 1953, 1954) Here

$$\gamma_2(x) \;=\; \int_0^\infty dy\ g_2\left(\sqrt{x^2 + y^2}\right) \tag{4.7}$$

is the two-point correlation function projected onto the sky,

$$\varepsilon_1 = 1 - \int_0^\infty \frac{dr\ r^2\ \theta_c^{\,2}(M)}{3} \Bigg/ \int_0^\infty \frac{dr\ r^2\ \theta_c^{\,2}(M)}{3} \tag{4.8}$$

is a measure of the "spread" of the luminosity functions and

$$\phi(z) \equiv \frac{q_o z + [q_o - 1]\ (\sqrt{1+2\ q_o z} - 1)}{q_o^{\,2}\ (1+z)^2} \tag{4.9}$$

an abbreviation. For power law two-point correlations, $g_2 = \xi_o / r^{\alpha}$, one obtains

$$G_2(1,2) = w_o/\theta^{\delta} - \frac{\varepsilon_1}{\mathcal{N}}\ \delta(X_1-X_2) \tag{4.10}$$

with $\delta = \alpha - 1$ and

$$w_o = \xi_o \frac{\Gamma(\frac{1}{2})\ \Gamma(\frac{\alpha-1}{2})}{\Gamma(\frac{\alpha}{2})}\ \frac{n_o^{\,2}}{\mathcal{N}^2} \int \frac{dr(1+z)^{\alpha}\ \theta_c^{\,2}\ r^{5-\alpha}}{\mathcal{G}^{6-\alpha}} \tag{4.11}$$

Provided the general-relativistic corrections can be neglected in (4.11), the two-point correlation functions $w_1(\theta)$ and $w_2(\theta)$ corresponding to plates with different limiting magnitudes m_1 and m_2 but taken for the same angular lag θ simply scale as

$$w_2(\theta) = f\frac{z_1}{z_2}\ w_1\frac{(z_2\ \theta)}{z_1} \tag{4.12}$$

(with f = 1), where z_1 and z_2 are characteristic redshifts corresponding to the depths of the galaxy samples. For a power law the amplitudes scale as $(z_1/z_2)^{\delta+1}$ at a given angular separation θ. However, (4.12) holds only in the Euclidean approximation and effects resulting from the expansion and space curvature are neglected. For galaxy surveys extending to magnitudes as faint as those in the Shane-Wirtanen catalog, a general-relativistic correction factor $f \neq 1$ must be introduced. The amount of the correction can be minimized, if, instead of the redshift, some suitable "effective distance" is introduced in (4.12) (Groth and Peebles 1977). Otherwise, fairly large values of f (of the order $\sim$4 for a limiting redshift of m = 20 mag, cf. Fig. 5 in Dautcourt 1977) must be expected. f is nearly independent of the deceleration parameter q_0, but depends to some degree on the luminosity function of galaxies and on the assumed K correction. The effect should increase the clustering amplitudes of very faint galaxies compared to those expected in a static Euclidean universe. Surprisingly, it appears that the observed degree of clustering does not show this relative increase but behaves like in a Euclidean universe. A possible interpretation is that the space clustering of galaxies increases on cosmic time scales (see section 4.2).

4.2. Observational data. Interpretation

A number of galaxy catalogs have been used to determine the surface correlation functions from observations. Table 3 collects some of the data obtained for the two-point function. The "limiting magnitudes" are derived from the mean number density of galaxies using relation (4.1). The resulting m_{lim} depend on the luminosity function and on the adopted K correction. We have used two luminosity functions, one in the form given by Peebles and Hauser (1974):

$$\Phi(M) = \begin{cases} a \cdot 10^{\beta(M - M^*)} & \dots \quad M - M^* \geq 0 \\ a \cdot 10^{\alpha(M - M^*)} & \dots \quad M_0 - M^* < M - M^* < 0 \\ 0 & \dots \quad M < M_0 \end{cases}$$

with $M_0 = M^* + (\lg \varphi_0)/\alpha$, and with the standard values $\alpha = 0.75$, $\beta = 0.25$, $\varphi_0 = 0.01$ (a a normalizing factor). The characteristic magnitude M^* was assumed to be $M^* = -19^m.5 + 5 \lg h_0$ ($h_0 = H_0/100$, H_0 the Hubble constant in usual units), whereas Peebles has $M^* = -18.6$ for $h_0 = 1$.

The second luminosity function was Kiang's (1961):

Table 3. Galaxy surveys for estimates of the two-point correlation function

Galaxy survey and field location	area	$\langle\bar{N}\rangle$ gal/□°	m_{lim} (Peebles' LF)	m_{lim} (Kiang's LF)	k_1	w_o	range in θ	range in r (Mpc)	evolution index s Peebles' LF	evolution index s Kiang's LF
Zwicky et al. 1961-67 $b^{II} > 40^o$, $\delta \geqslant 0^o$	6000	.626	15^m	15^m		14.9	$.2^o$-30^o	0.4-56	-	-
Shane and Wirtanen 1967 $b^{II} > 40^o$	7400	45	18.7	18.5	3.1	1.32	1^o-10^o	8-80	6.6	6.9
Rudnicki et al. 1973 $\alpha=11^h19^m$, $\delta=35^o35'$	38	334	20.85	20.45	3.2	.386	7'5-1^o	1.8-14	7.6	7.7
Dautcourt et al. 1978										
I. $\alpha=13^h34^m.3$, $\delta=30^o.1$	9.8	590	21.5	21.1	3.2	.206			10.2	9.8
II. $\alpha=13^h35^m$, $\delta=27^o.5$	9.7	801	21.9	21.4	3.2	.154	5'-1^o	1.8-22	11.1	10.4
III. $\alpha=13^h40^m$, $\delta=30^o.1$	9.3	549	21.5	21.0	3.2	.296			7.4	7.6
IV. $\alpha=13^h49^m$, $\delta=27^o.2$	9.8	786	21.9	21.4	3.2	.286			5.8	6.4
Dodd et al. 1975 $\alpha=2^h49^m.3$, $\delta=-32^o.2$	1.96	1480	22.4	22.3	1.7	.139	.25'-10'	.1-3.2	7.2	8.5

$$\Theta(M) = \begin{cases} a(M-M_0)^3 & \dots\ 0 < M-M_0 \leq 2.5 \\ (2.5)^3\, a\ 10^{0.2(M-M_0)-0.5} & \dots\ 2.5 < M-M_0 < 8 \\ 0 & \dots\ \text{remaining values of M,} \end{cases}$$

here $M_0 = -21^m.9 + 5\lg h_0$. For the K correction a term linear in z, $K = k_1 z$, was assumed. k_1 was calculated by averaging the contribution from different galaxy types weighted with the space distribution of types.

With (4.11) one calculates the space amplitude ξ_0 of the two-point function S(r) given in Table 4. Differences with Peebles arise mainly from our use of $M^* = -19.5$ in the luminosity function (the most recent value derived by Peebles and collaborators is $h_0 r_0 = 4.6$ Mpc in $\xi = (r_0/r)^{1.77}$).

Table 4. Galaxy number density n_0 and amplitude ξ_0 of the space two-point correlation function determined from the Zwicky catalogs for two luminosity functions

	$n_0(\mathrm{Mpc}^{-3})$	ξ_0	$h_0 r_0$(Mpc)
Kiang's LF	0.023	60	5.6
Peebles' LF	0.016	114	8.1

The empirical estimates of the two-point correlation functions show expected and unexpected features. First, at a given angular scale the clustering amplitudes measured by w_0 decreases with increasing depth of the survey, as expected from equation (4.12). Secondly, as already noted in section 4.1, this decrease of clustering degree occurs more rapidly than expected, if "no evolution" of the galaxy distribution takes place. Indeed, if the general-relativistic formula (4.11) is employed, the calculated values w_0 are considerably larger than the observed ones. To make clear what evolution of galaxy clustering means, let us assume that an astronomer measures the distribution of galaxies around himself using his local radial coordinate x. He may find some power law $\xi = \tilde{\xi}_0/x^{\alpha}$ for the two-point correlation function. If $\tilde{\xi}_0$ is independent of the cosmic time t, the astronomer would speak about "no evolution". Any change in the galaxy distribution may be represented by $\xi_0 = \tilde{\xi}_0(1+z)^{-s}$ (with $\tilde{\xi}_0$ independent

of cosmic time), an "evolution index" $s > 0$ indicating a slow increase of degree of clustering on cosmic time scales. With respect to a comoving radial coordinate $r = x(1 + z)$ we have

$$\xi(r) = \frac{\xi_0 (1 + z)^{\alpha - s}}{r^{\alpha}} \qquad (4.13)$$

w_0-values were calculated from (4.11) with (4.13) and with s adjusted so that the observed value of w_0 agrees with the theoretical one. The resulting s shown in Table 3, column 10 and 11, are fairly large. Large values of s also result from deep galaxy samples discussed by Phillipps et al (1978), which are not included in Table 3. Taken at face value the data would indicate that the clustering amplitudes increase in cosmic times.

Certain evolution effects in the observed sense should be present because the system of galaxies is unstable gravitationally. The density contrast $\langle \delta n^2 \rangle / n^2$ is known to increase like $\sim t^{4/3} = (1+z)^{-2}$ in an Einstein-deSitter universe (Zeldovich and Novikov 1975), and the same increase applies to the correlation function $\xi(r)$ for sufficiently large spacings r in the fluid limit of the kinetic equations (Inagaki 1976). Thus one expects $s = 2$ for large r. For smaller r, particularly if the discrete nature of the interacting galaxies comes into play, the picture is more complicated (Fall and Saslaw 1976, Inagaki 1976, Yahil 1976, Davis and Peebles 1977, Press and Lightman 1978, Silk and White 1978). It is not clear, however, if the large values of s found here could be accounted for by gravitational interactions. There are a number of other effects which tend to produce a large s (Dautcourt and Richter 1978). Errors in the identification of galaxies will increase s considerably. On the other hand, ordinary galactic luminosity evolution should have no essential influence on the data in Table 3, since the catalogs do not extend to extremely faint magnitudes. According to Tinsley (1977), however, contamination of the counts by very bright and highly redshifted young galaxies could influence the observed clustering degree. A detailed study of the galaxy clustering extending to depths now in the range of the Soviet 6 m telescope (Karachentsev and Kopylov 1977) should help to answer several of these questions.

A further property of the empirical two-point correlation function is an increase of the slope δ at angular distances $\theta > \theta^*_{SW} \simeq 2^\circ.5$ in the Shane-Wirtanen catalog (Davis et al. 1977, Groth and Peebles 1977, Dautcourt 1978). One usually assumes that this steepening is also present in the space correlation function at a linear separation

of $h_o r^* \simeq 9$ Mpc. However, the observed behaviour of the surface function $w(\theta)$ could also be produced by a single power law $\xi \sim r^{-(\delta+1)}$ with a cutoff at some $r \gg r^*$. An interesting interpretation of the break results from the fact that r^* is only slightly larger than r_o (defined by $\xi(r_o) = 1$). Hydrodynamically, the scale r_o corresponds to the transition from the linear treatment of gravitational density perturbations to the non-linear regime (for $r < r_o$). For $r < r_o$ the density perturbations become rapidly larger compared with the mean density, and bound systems fragment out of the general distribution of galaxies. In the kinetic treatment r_o signifies the inset of strong turbulence with higher-order correlation functions becoming large.

Empirical estimates of the three-point (Groth and Peebles 1977, Peebles and Groth 1975, Peebles 1975) and four-point (Fry and Peebles 1978) correlation functions (mainly from the Zwicky and Shane-Wirtanen catalogs) support this picture. In view of the expected homogeneity and isotropy of the space random process one may assume that the surface correlation functions G_3 and G_4 depend only on the mutual angular distances of three or four points on the sky (see also the Appendix):

$$\begin{aligned} G_3 &= G_3(\theta_{23}, \theta_{13}, \theta_{12}), \\ G_4 &= G_4(\theta_{23}, \theta_{13}, \theta_{12}, \theta_{14}, \theta_{24}, \theta_{34}). \end{aligned} \tag{4.14}$$

Moreover, the empirical data show that G_3 and G_4 can be represented in terms of products of two-point functions:

$$\begin{aligned} G_3 &= Q(w_{23}w_{13} + w_{13}w_{12} + w_{23}w_{12}) \\ G_4 &= R_a(w_{12}w_{23}w_{34} + (\text{further 11 terms})) \\ &\quad + R_b(w_{12}w_{13}w_{14} + (\text{further 3 terms})), \end{aligned} \tag{4.15}$$

where we have used the notation $w_{12} = w(\theta_{12})$ etc. Similar equations hold for the space functions:

$$\begin{aligned} g_3 &= q(\xi_{23}\xi_{13} + \xi_{13}\xi_{12} + \xi_{23}\xi_{12}), \\ g_4 &= r_a(\xi_{12}\xi_{23}\xi_{34} + (\text{sym})) + r_b(\xi_{12}\xi_{13}\xi_{14} + (\text{sym})), \end{aligned} \tag{4.16}$$

where the parameters q, r_a, r_b are connected with their surface counterparts Q, R_a, R_b by integral relations similar to (4.11). Numerical estimates are given in Table 5. The fact that q, r_a and r_b are of order 1 shows that the higher-order correlation functions dominate for small space separations r with $\xi(r) > 1$. In this region of "strong clustering" a representation of the n-point probability distribution

function f(1,...N) in terms of low-order correlation functions (Mayer cluster expansion, equation (2.11)) appears to be not practical. Instead one may turn to other measures of galaxy clustering like those discussed in section 3.

Table 5. Coefficients for the three-point and four-point correlation functions (Q for the Shane-Wirtanen catalog) (Groth and Peebles 1977, Fry and Peebles 1978)

Q	q	r_a	r_b
1.56±0.22	1.29±0.21	2.5±0.6	4.3±1.2

5. Concluding remarks

The approach given here has considered the correlation functions as basic physical quantities. In view of the physical processes involved in the formation of galaxies and clusters of galaxies this might not necessarily be the best starting point. One should remember the theory of galaxy distribution developed by J. Neyman, E.L. Scott and their collaborators (Neyman et al. 1956) where the existence of clusters in space with randomly distributed centers and with definite structure properties was the basic assumption. This model gives rise to a well-defined sequence of two-point, three-point and higher order correlation functions, which could be calculated using, e.g., a method recently elaborated by McClelland and Silk (1977 a,b). If one considers models for the formation of galaxies such as the model of adiabatic perturbations developing into "pancakes" by nonlinear interactions arranged in a kind of honeycomb structure (Zeldovich 1978), also specific correlation functions for the galaxy distribution must be expected. If, on the other hand, galaxy clustering builds up from small units (Press and Schechter, 1974) the resulting properties of the galaxy correlation function are expected to be quite different. In principle, this could allow us to distinguish between different models of the origin of galaxies and galaxy clusters.

Appendix: Notation for cell moments

A surface correlation function G_r is usually written $G_r(X_1 \ldots X_r)$ as a function of the arguments, where X_1 denote the two angular coor-

dinates at the sky. If no misunderstanding is possible we simply write $G_r(1,\ldots r)$. Partial integration over a cell with respect to some coordinates is denoted by

$$G_{ri}(X_1,X_2\ldots X_i) = \int \frac{dY_1\ldots dY_{r-i}}{\omega^{r-i}} G_r(X_1\ldots X_i, Y_1\ldots Y_{r-i}). \qquad (A\ 1)$$

This includes for i = 0 the cell moment of r-th order:

$$G_{ro} = \int \frac{dY_1\ldots dY_r}{\omega^r} G_r(X_1\ldots X_r) \qquad (A\ 2)$$

To distinguish cells of different shape or size, we may write $G_{ro}^{(1)}$, $G_{ro}^{(2)}$ etc. In some cases, where the integration is not taken over the same cell, we use a similar notation, for instance

$$G_{20}^{(1,2)} = \int_{\omega_1}\int_{\omega_2} G_2(X_1,X_2)\ \frac{dX_1 dX_2}{\omega_1\,\omega_2} \qquad (A\ 3)$$

where $\omega_1 \neq \omega_2$.

The notations (A 1) - (A 3) apply to the general case where no symmetry property for the correlation functions was assumed. If we assume homogeneity and isotropy for the surface random process, it follows that $F_1 = 1$, $G_2(X_1,X_2)$ (in general a function of four independent arguments) depends only on the angular distance Θ_{12} between the two points X_1 and X_2, $G_3(X_1,X_2,X_3)$ depends only on the angular distances Θ_{12}, Θ_{13}, Θ_{23} between the three points X_1, X_2 and X_3 etc. This simplification was used in (4.6 - 4.7).

References

Abell, G.O., 1958, ApJ. Suppl. 3, 211
Abell, G.O., 1974, in M.S. Longair (ed.), Confrontation of cosmological theories with observational data, D. Reidel P.C.,Dordrecht
Bok, B.J., 1934, Harvard Bull. No. 895
Dautcourt, G., 1977, Astronom. Nachr. 298, 253
Dautcourt, G., Richter, N., 1978, Astronom. Nachr. 299, 171
Dautcourt, G., Kempe, K., Richter, L. and Richter, N., 1978, Astronom. Nachr. 299, 177
Dautcourt, G., 1978, in press
Davis, M. and Peebles, P.J.E., 1977, ApJ. Suppl. 34, 425
DeVaucouleurs, G., 1971, Publ. Astron. Soc. Pacific 83, 113
Dodd, R.J., Morgan, D.H., Nandy, K., Reddish, V.C. and Seddon, H., 1975, Montly Notices R.A.S. 171, 329
Doroshkevich, A.G. and Zeldovich, Ya.B., 1975, Astrophys. and Space Science 35, 55
Fall, S.M., Saslaw, W.C., 1976, ApJ. 204, 631
Fall, S.M., Geller, M.J., Jones, B.J.T., White, S.D.M., 1976, ApJ 205, L 121

Flin, P., Machalski, S., Maslowski, J., Urbanik, M., Zieba, A., Zieba, S., 1974, in M.S. Longair (ed.), Confrontation of cosmological theories with observational data, D. Reidel P.C., Dordrecht
Flin, P., 1974, Memorie della Societa Astronomica Italiana 45, No. 3/4
Flin, P., 1977, Acta Cosmologica 6, 19
Gilbert, P.H., 1971, Astrophys. Space Sci. 14, 3
Gott, S.R., Rees, M.D., 1975, Astron. Astrophys. 45, 365
Fry, J.N. and Peebles, P.J.E., 1978, ApJ 221, 19
Groth, E.J. and Peebles, P.J.E., 1977, ApJ 217, 385
Hubble, E., 1934, ApJ 79, 8
Inagaki, S., 1976, Publ. Astr. Soc. Jap. 28, 77
Kalinkov, M., 1972, Proceed. First European Astron. Meeting 3, 142
Kalinkov, M., 1978, in press
Karachentsev, L.D. and Kopylov, A.I., 1977, Pisma Astron. Zh. 3, 246
Klimontovich, Yu.L., 1964, Statistical Theory of nonequilibrium processes in a plasma (in Russian), Isdat. MGU, Moscow
Klimontovich, Yu,L., 1975, Kinetic theory of nonidial gas and nonidial plasma (in Russian), Nauka, Moscow
Layzer, D., 1956, A.S. 61, 383
Layzer, D., 1959, Galaxy clustering its description its interpretation, preprint
Limber, D.N., 1953, ApJ 117, 134
Limber, D.N., 1954, ApJ 119, 655
Mayer, S.E. and Mayer, M.G., 1940, Statistical Mechanics of Fluids, New York
McClelland, J. and Silk, J., 1977, ApJ. 216, 665
ApJ. 217, 331
Montgomery, D.C. and Tidman, D.A., 1964, Plasma Kinetic Theory, McGraw-Hill, New York
Neyman, S., Scott, E,L. and Shane, C.D., 1956, Proc. Third Berkeley Symp. Math. Stat. and Prob.
Peebles, P.J.E. and Hauser, M.G., 1974, ApJ. Suppl. 28, 19
Peebles, P.J.E., 1975, ApJ. 196, 647
Peebles, P.J.E. and Groth, E.J., 1975, ApJ. 196, 1
Phillips, S., Fony, R., Ellis, R.S., Fall, S.M. and MacGillivray, H.T., 1978, Mon. Not. R. astr. Soc. 182, 673
Press, W.H., Schechter, P., 1974, ApJ. 187, 425
Press, W.H., Lightman, A.P., 1978, ApJ. 219, L 73
Rudnicki, K., Dworak, T.Z., Flin, P., Baranowski, B. and Sendrakowski, A., 1973, Acta Cosmologica 1
Rudnicki, K., Zieba, S., 1978, in M.S. Longair and J. Einasto (ed.), The Large Scale Structure of the Universe, Reidel, P.C., Dordrecht
Saslaw, W.C., 1972, ApJ. 177, 17
Seldner, M., Siebers, B., Groth, E.J., Peebles, P.J.E., 1977, A.J. 82, 249
Silk, J., White, S.D., 1978, ApJ. 223, L 59
Shane, C.D. and Wirtanen, C.A., 1967, Publ. Lich Obs. 22, 1
Soneira, R.M. and Peebles, P.J.E., 1977, ApJ. 211, 1
Turner, E.L. and Gott, J.R., 1975, ApJ. 197, L 89,
Wagoner, R.V., 1967, Nature 214, 766
Wagoner, R.V. and Bogart, R.S., 1973, ApJ. 181, 609
Yahil, A., 1976, ApJ. 204, L 59
Zeldovich, Ya.B., Novikov, I.D., 1975, Structure and evolution of the universe (in Russian) Nauka, Moskow
Zeldovich, Ya.B., 1978, in M.S. Longair, S. Einasto (ed.), The Large-Scale Structure of the Universe, Reidel, Dordrecht
Zieba, A., 1975, Acta Cosmologica 3, 75
Zwicky, F., 1957, Morphological Astronomy, Springer, Berlin

Zwicky, F., Herzog, E., Wild, P., Karpowicz, M. and Kowal, C.T., 1961-1968, Catalogue of Galaxies and Clusters of Galaxies, I-VI, Pasadena, California Institute of Technology.

QUESTIONS TO INFALLIBLE ORACLE

M. Heller

1. Introduction: purely terminological question

Within the set of all spatially-homogeneous solutions of Einstein's field equations there is a zero-measure sub-set of solutions which are both spatially-homogeneous and isotropic, and which are called Friedman, Lemaître, Robertson-Walker, and all possible combinations of these names (with the restriction that Robertson and Walker almost always go together), cosmological models. Works of Robertson and Walker are chronologically later and deal rather with symmetric spaces from the purely mathematical point of view than with the relativistic world models. It remains, however, the question of priority as far as contributions of Friedman and Lemaître are concerned. Belonging to the zero-measure set of people who have read both cosmological papers of Friedman and almost all papers of Lemaître (who is the author of more than one hundred papers); I shall deal with this question.

In the Institute of Geophysics and Astronomy (bearing now the name of Georges Lemaître) at the Catholic University of Louvain (now: Louvain-la-Neuve) there is an archive of papers, notes, letters, etc. left by Lemaître, who was there a professor all his scientific life. During my two six months visits in Louvain-la-Neuve I spent many hours, together with Professor Odon Godart, former assistant of Lemaître, in close contact with Lemaître's papers. I will present here some of our findings, throwing a beam of light onto the early history of the relativistic cosmology.

2. The works of Friedman

Before fundamental works of Friedman and Lemaître, two cosmological models were known: original Einstein's model (1917) containing uniformly distributed dust-like matter, and de Sitter's model (1917) containing no matter. Both models were believed to be static. It was Lemaître (1925) who - by introducing a suitable coordinates - has discovered stationary (but non-static) character of the de Sitter solution.

Two Friedman's works (1922, 1924) chronologically preceded those

of Lemaître. However, both articles of Friedman, although published in the German journal "Zeitschrift für Physik", remained unknown for a long time to the wider scientific public.

The first Friedman's work, entitled "Über die Krümmung des Raumes" appeared in 1922. At the beginning the author carefully enumerates all assumptions necessary to construct a cosmological model. He retains two Einstein's postulates: the cosmological constant and a constant, positive curvature of space; he abandons, however, Einstein's presupposition that the universe has to be static. This leads immediately to the so-called to-day Friedman's equations (with $k = +1$), governing the overall evolution of the universe. Evolution is admitted by Friedman as a mathematical possibility. The author gives explicitly the reason for such a restriction: this is because of lack of empirical data "which could give any estimates and provide an answer to the question which space-time corresponds to our universe".

The work of Friedman is mathematically very elegant. It gives the full discussion of all possible solutions (without, however, writing them down explicitly), within the considered class of models. This discussion was afterwards repeated in the famous monograph by Tolman (1934), and often mistakenly attributed to Robertson (1933).

The second work of Friedman, which appeared in 1924, is entitled "Über die Möglichkeit einer Welt mit konstanter negativer Krümmung des Raumes". The "Friedman equation" is deduced for the case of constant, negative curvature of space, and the proof is given that this equation admits non-stationary solutions with positive matter density (without giving their discussion as in the first paper).

Motivation of the second paper is more philosophical. Friedman wants to contribute to the old problem of whether the universe is spatially finite or infinite. He treats this problem in a very modern style, showing that the answer depends not only on the field equations themselves but also on the assumed global topology of space-time. (For the recent treatment of this problem see the beautiful paper by Ellis (1971)).

It is rather curious that Friedman did not consider the class of world models with the vanishing spatial curvature. This neglect was afterwards made up by Robertson (1933).

After the first work of Friedman, Einstein (1922) published a short note claiming to find an error in Friedman's computations. In the second note (1924), which was an answer to a private letter of Friedman, he admitted that it was himself who comitted an error. Einstein did not like an idea of the expanding universe. He fully appre-

ciated the value of Friedman's work in 1931, six years after the death of the Russian cosmologist.

3. The universe of increasing radius

The fundamental work of Lemaître was published in 1927, in a local Belgian journal "Annales de la Société Scientifique de Bruxelles", and also remained unnoticed by broad scientific public. Eddington, working on the problem of instability of Einstein's static model, came across the Lemaître paper, appreciated its value, and found in it (implicite) the solution of his own problem. Later on, Eddington managed to publish the English translation of Lemaître's work in the "Monthly Notices of the Royal Astronomical Society" (1931).

The idea of the solution came to Lemaître when listening to Hubble's lecture in United States (private information from Professor Godart), and from the beginning this idea was strictly connected with observational investigations of the universe. Therefore the title of Lemaître's paper "A Homogeneous Universe of Constant Mass and Increasing Radius Accounting for the Radial Velocity of Extra-galactic Nebulae" is significant.

Lemaître's intention was to find a model intermediate between Einstein static universe and empty de Sitter world. After the work on new coordinates for the de Sitter universe, he knew that the model looked for has to be non-stationary, "with increasing radius". This led Lemaître to the Friedman equation for the constant positive space curvature. Wanting to have something between Einstein and de Sitter models, Lemaître adjusted constants of integration to their values in Einstein's static universe, and obtained rather special solution, known afterwards under the name of Eddington-Lemaître world model, that describes the expanding universe with non-vanishing matter density, which - as time goes to minus infinity - approaches asymptotically the static Einstein universe.

Friedman's equation in Lemaître's paper incorporates a term corresponding to pressure of matter, however, if model is compared with astronomical observations, Lemaître puts this term equal zero. The aim of Lemaître is to study the real universe and not a mathematical structure of the Friedman equation. For this reason he does not consider other possible solutions, which - in his opinion - provide too short time scale as compared to that of stellar evolution (the minimum of the radius R "would generally occur at time of the order of R_o, say 10^9 years - i.e. quite recently for stellar evolution"). Lemaître de-

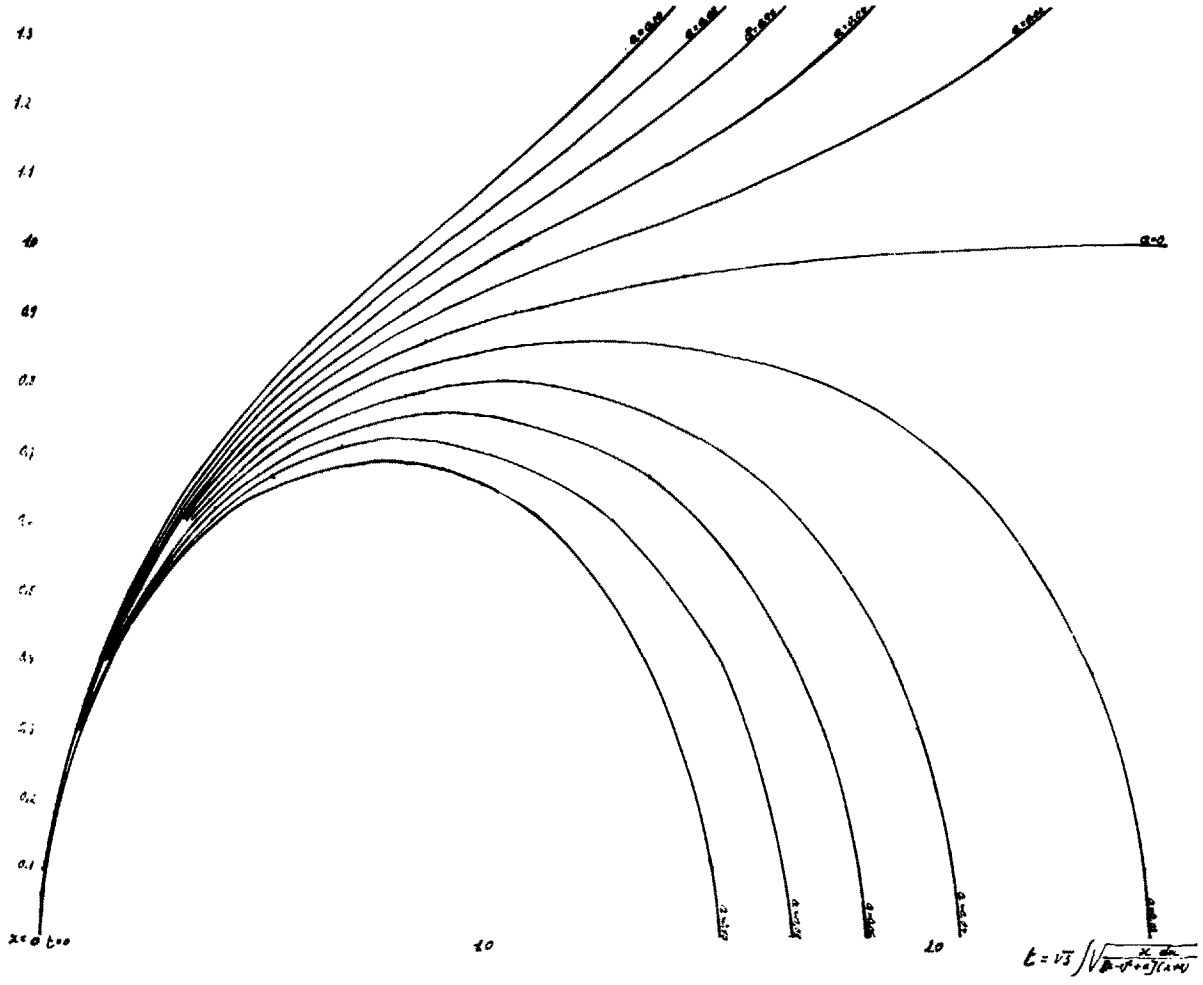

Fig. 1. Original drawing of Lemaître showing all possible solutions of Friedman equations with constant positive space curvature. (Courtesy of Professor Odon Godart).

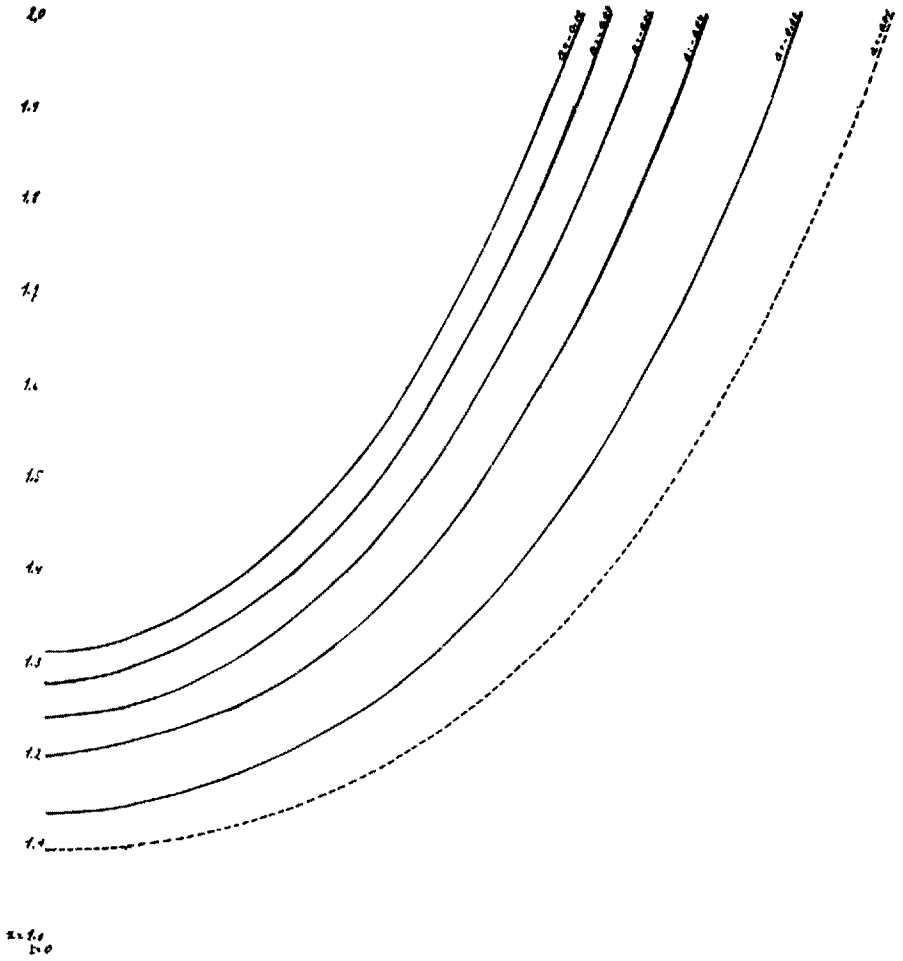

Fig. 2. Original drawing of Lemaître showing all possible solutions of Friedman equations, solution adopted by Lemaître indicated with dashed line. (Courtesy of Professor Odon Godart) .

rives from his model the formula for the Doppler effect, for the case "when the light source is near enough" he obtains the "Hubble law" (which was published by Hubble two years later), and compares it with the 43 red-shift measurements made by Strömberg and 42 made by Hubble himself. The general conclusion is: "The receding velocities of extra-galactic nebulae are a cosmic effect of the expansion of the universe".

In Lemaître's archive in Louvain-la-Neuve, we have found a red pad with the inscription "1927", containing: the proofs of Lemaître's 1927 paper, some notes in handwriting connected with this paper, and two diagrams (see Fig. 1 and 2) presenting all possible solutions for the case of constant positive space curvature with the solution adopted by Lemaître distinguished with the dashed line. This is interesting from the historical point of view, since people believe that Lemaître learned about the existance of other solutions much later (about 1931). If we remember that Friedman gave only the qualitative discussion of the solutions, it is highly probable that Figs. 1 and 2 present, for the first time in the history, time evolution of "closed universes", reproduced afterwards almost in every textbook of cosmology.

4. Quantum cosmology

Lemaître was not happy with his results. On the other hand, Eddington who wrote that "the notion of a beginning of the present order of Nature is repugnant to me", widely propagated Lemaître's model with minus time infinity. The quoted sentence of Eddington comes from his presidential address to the Mathematical Association, published in Nature and entitled: "The End of the World: from the Standpoint of Mathematical Physics" (1931). About three weeks later (!) in the same journal a short note of Lemaître appeared, which clearly was inspired by Eddington's address. Lemaître's note bears the title: "The Beginning of the World from the Point of View of Quantum Theory" (1931).

> This is a very short note, and I shall quote its larger parts: "Sir Arthur Eddington states that, philosophically, the notion of a beginning of the present order of Nature is repugnant to him. I would rather be inclined to think that the present state of quantum theory suggests a beginning of the world very different from the present order of Nature. Thermodynamical principles from the point of view of quantum theory may be stated as follows: (1) Energy of constant total amount is distributed in discrete quanta. (2) The number of distinct quanta is ever increasing. If we go back in the course of time we must find fewer and

fewer quanta, until we find all the energy of the universe packed in a few or even in a unique quantum.

Now, in a atomic processes, the notion of space and time are no more than statistical notions; they fade out when applied to individual phenomena involving but a small number of quanta. If the world has begun with a single quantum, the notions of space and time would altogether fail to have any meaning at the beginning; they would only begin to have a sensible meaning when the original quantum had been divided into a sufficient number of quanta. (...)

Clearly the initial quantum could not conceal in itself the whole course of evolution; but according to the principle of indeterminacy, that is not necessary. Our world is now understood to be a world where something really happens; the whole story of the world need not have been written down in the first quantum like a song on the disc of a phonograph. The whole matter of the world must have been present at the beginning, but the story it has to tell may be written step by step."

These are the beginnings of the so-called Primeval Atom Hypothesis. The quantum considerations should be regarded as its first source.

5. Initial singularity

The second source of the Primeval Atom Hypothesis were Lemaître's considerations on the problem of the initial singularity. Lemaître discussed this problem with Einstein who did not like the idea of a "beginning". Hoping that the initial singularity appears as a "by--product" of the symmetry assumptions, Einstein suggested to Lemaître to consider a simple anisotropic metric (called today Bianchi I). Lemaître (1933) easily wrote down the corresponding field equations, and - using an equality which essentially is the so-called energy dominance condition of the contemporary Hawking-Penrose singularity theorems - succeeded to demonstrate that in this particular case the singularity is present.

Lemaître emphasises that this is not a formal proof, that the singularity cannot be removed by introducing anisotropy (because the considered metric is not the most general possible) but nevertheless it does suggest that even in the more general case the anisotropy is not an effective mechanism for removing singularities because "it acts in an opposite direction".

A simple physical consequence of the inevitability of the ini-

tial singularity is the fact that - if we go back in time - matter "should have higher and higher temperature, much higher than the critical temperature of fluids, and nothing prevents it to reach the degree of compression comparable to the interior of Sirius' companion".

In this way the concept of the initial singularity is a geometric support of the physical idea of the Primeval Atom.

6. Further developments: the background radiation

The Primeval Atom Hypothesis consists of further developments of the above premisses. The idea was improved and evolved in many works of Lemaître. We cannot go here into details. Instead, we shall quote a longer passage from one of Lemaître's later papers:

> "The physical beginning which fits the solution of Friedman's equation starting from R = 0 is provided by the Primeval Atom Hypothesis.
>
> Here the word "Atom" should be understood in the primitive Greek sense of the word. It is intended to mean absolute simplicity, excluding any multiplicity. The Atom is so simple that nothing can be said about it and no question raised. It provides a beginning which is entirely inaccessible.
>
> It is only when it has split up into a large number of fragments by filling up a space of small, but non strictly zero radius, that physical notions begin to aquire some meaning.
>
> The first question which has to be considered is whether the resulting assembly of particles has to be described as a gas.
>
> If one gives an affirmative answer to this question, one has to face the difficulty of understanding how such a gas, which presumably filled up an expanding space, has to be able, later on, to divide itself into separate nebulae.
>
> To be more precise, we must make clear what has to be considered as characterizing a gas. It is not enough to have an assembly of a large number of particles. In order to be called a gas, such an assembly must have velocities with a distribution that is strongly concentrated around a mean velocity, the velocity of the gas, and distributed around this velocity according to a law not too different from the Maxwellian distribution which is realized in ordinary gases.
>
> On the other hand, a mere assembly of particles with velocities spreading in every direction with speeds of the same

order of magnitude could not be considered as a gas. It should be described as an assembly of corpuscular rays, as corpuscular radiation.

It is true that, by collisions, such radiation would finally reach a state of statistical equilibrium and become a gas. But in the extreme condition of expansion, starting (theoretically) with infinite velocity, it is not likely that such a statistical equilibrium would have had time to establish itself.

From that point of view, the problem which cosmology has to face is to understand how gas would finally arise from the primeval radiation and then organize itself into nebulae and secondly to understand what would arise from the part of this primeval radiation which would have escaped condensation into gases.

The second point gives an interpretation of the observed cosmic radiation, which may, of course, be only a partial one. In discussing this aspect of the theory, one must take into account for the rays the reduction of intensity due to the expansion. This phenomenon, quite analogous to the red shift of light, reduces the intensity of the rays in proportion to 1/R. The total intensity of the cosmic rays is about 1/10 000 of the total energy of matter condensed in the stars. This means that cosmic rays and matter would have been of the same order of magnitude when the radius was only one ten-thousandth of its present value." (Lemaître (1958)).

Here we have, if not a simplified version of a contemporary standard model of the universe, at least a clearly formulated programme for such a model. With one exception: cosmic rays - Lemaître's candidate for the relic primeval radiation should be replaced by the microwave background radiation discovered in 1965. Two weeks before his death Lemaître learned about this discovery (private communication of O. Godart) and was happy that his hypothesis has acquired an experimental support.

7. Instead of conclusions

From the contemporary perspective we may look on Lemaître's work as on the initiation of the "physical cosmology" (not only a geometrical frame for the structure-evolution of the universe, as it was in earlier cosmological investigations). However, during his life Lemaître met sometimes scepticism and even ironical atmosphere.

Among letters of Lemaître we have found the following post card,

l'Abbé Lg. Lemaître,
l'Université,
Louvain
Belgium.

Cambridge.
17-4-34.

congratulations from the Irish on the remunerative creation of the Universe

P. Kapitza M. Born. Ch. T. Wilson
P. A. M. Dirac R. Fraser A. Compton
E. T. S. Walton P. C. Vaux W. D. Lewis.
J. D. Cockcroft. H. A. Wooster.

Fig. 3. Copy of a post card send to Lemaître from Cambridge on April 17-th, 1934. (Courtesy of Professor Odon Godart).

mailed to him from Cambridge: "Cambridge, 17.4.1934. - Congratulations from the club on the remuneration creation of the universe." Among eleven signatures there are names of P. Kapitza, P.A.M. Dirac, M. Born. After Lemaître's death Dirac (1968) wrote a beautiful review article about Lemaître's contributions to the modern science in which he calls Lemaître the greatest cosmologist of our times.

In 1931 the British Association organized discussion on "the question of the relation of the physical universe to life and mind". Contributions to this discussion (by J. Jeans, G. Lemaître, W. de Sitter, A. Eddington, R.A. Millikan, E.A. Milne, General J.C. Smuts, Bishop E.W. Barnes, and O. Lodge), under the common title "The Evolution of the Universe", from a Supplement to Nature (October 24, 1931). Even now this discussion is very interesting to read. To-day we know more "cosmic mechanisms" and technicalities about the universe but we are not much closer to the most fundamental answers.

At the end of his contribution Sir James Jeans said:

"Suppose some infallible oracle affered to give a "Yes" or "No" answer to two scientific questions for each of us. Personally, I think I might choose as my two questions:

1. Does the main energy of stellar radiation come from the annihilation of matter?

2. Is the universe expanding at about the rate indicated by the spectra of the nebulae?"

Lemaître was a second speaker, he assumed Jeans' style:

"If I had to ask a question of the infallible oracle alluded to by Sir James Jeans, I think I should choose this: "Has the universe ever been at rest, or did the expansion start from the beginning?" But, I think, I would ask the oracle not to give the answer, in order that a subsequent generation would not be deprived of the pleasure of searching for and of finding the solution."

Owing to Lemaître's generosity we can investigate to-day mysteries of the universe and leave many unanswered questions for future generations.

References

de Sitter, W., 1917, Proc. Akad. Wetench. Amsterdam, 19, 1217.
Dirac, P.A.M., 1968, The Scientific Work of Georges Lemaitre, in Pontificial Academiae Scienciarum Commentarii, vol.II, 1.
Eddington, A.S., 1931, Suppl. Nature, March 21.
Einstein, A., 1917, Sitzungsber. preuss. Akad. Wiss. p.142.
1922, Z. Phys. 11, 326.
1923, Z. Phys. 16, 228.

Ellis, G.F.R., 1971, Gen. Rel. Grav. 2, 7.
Friedman, A., 1922, Z. Phys. 10, 377.
1924, Z. Phys. 21, 326.
Lemaître, G., 1925, J. Math. and Phys. 4, 37.
1927, Ann. Soc. Sci. Bruxelles 47A, 29.
1931, Mon. Not. Roy. Astr. Soc. 91, 483.
1931, Nature, March 21.
1933, Ann. Soc. Sci. Bruxelles 53A, 51.
1958, in Semaine d'Etude sur le Probleme des Populations Stellaires, Pont. Acad. Sci.
The Evolution of the Universe, 1931, Suppl. Nature, October 24.

Selected Issues from

Lecture Notes in Mathematics

Vol. 594: Singular Perturbations and Boundary Layer Theory, Lyon 1976. Edited by C. M. Brauner, B. Gay, and J. Mathieu. VIII, 539 pages. 1977.

Vol. 596: K. Deimling, Ordinary Differential Equations in Banach Spacea. VI, 137 pages. 1977.

Vol. 605: Sario et al., Classification Theory of Riemannian Manifolds. XX, 498 pages. 1977.

Vol. 606: Mathematical Aspects of Finite Element Methods. Proceedings 1975. Edited by I. Galligani and E. Magenes. VI, 362 pages. 1977.

Vol. 607: M. Métivier, Reelle und Vektorwertige Quasimartingale und die Theorie der Stochastischen Integration. X, 310 Seiten. 1977.

Vol. 615: Turbulence Seminar, Proceedings 1976/77. Edited by P. Bernard and T. Ratiu. VI, 155 pages. 1977.

Vol. 618: I. I. Hirschman, Jr. and D. E. Hughes, Extreme Eigen Values of Toeplitz Operators. VI, 145 pages. 1977.

Vol. 623: I. Erdelyi and R. Lange, Spectral Decompositiona on Banach Spaces. VIII, 122 pages. 1977.

Vol. 828: H. J. Baues, Obstruction Theory on the Homotopy Classification of Maps. XII, 387 pages. 1977.

Vol. 629: W. A. Coppel, Dichotomies in Stability Theory. VI, 98 pages. 1978.

Vol. 630: Numerical Analysis, Proceedings, Biennial Conference, Dundee 1977. Edited by G. A. Watson. XII, 199 pages. 1978.

Vol. 636: Journées de Statistique des Processus Stochastiques, Grenoble 1977, Proceedings. Edité par Didier Dacunha-Castelle et Bernard Van Cutsem. VII, 202 pages. 1978.

Vol. 638: P. Shanahan, The Atiyah-Singer Index Theorem, An Introduction. V, 224 pages. 1978.

Vol. 648: Nonlinear Partial Differential Equations and Applications, Proceedings, Indiana 1976–1977. Edited by J. M. Chadam. VI, 206 pages. 1978.

Vol. 650: C*-Algebras and Applications to Physics. Proceedings 1977. Edited by R. V. Kadison. V, 192 pages. 1978.

Vol. 656: Probability Theory on Vector Spaces. Proceedings, 1977. Edited by A. Weron. VIII, 274 pages. 1978.

Vol. 662: Akin, The Metric Theory of Banach Manifolds. XIX, 306 pages. 1978.

Vol. 665: Journées d'Analyse Non Linéaire. Proceedings, 1977. Edité par P. Bénilan et J. Robert. VIII, 256 pages. 1978.

Vol. 667: J. Gilewicz, Approximants de Padé. XIV, 511 pages. 1978.

Vol. 668: The Structure of Attractors in Dynamical Systems. Proceedings, 1977. Edited by J. C. Martin, N. G. Markley and W. Perrizo. VI, 264 pages. 1978.

Vol. 675: J. Galambos and S. Kotz, Characterizations of Probability Distributions. VIII, 169 pages. 1978.

Vol. 676: Differential Geometrical Methods in Mathematical Physics II, Proceedings, 1977. Edited by K. Bleuler, H. R. Petry and A. Reetz. VI, 626 pages. 1978.

Vol. 678: D. Dacunha-Castelle, H. Heyer et B. Roynette. Ecole d'Eté de Probabilités de Saint-Flour. VII-1977. Edité par P. L. Hennequin. IX, 379 pages. 1978.

Vol. 679: Numerical Treatment of Differential Equations in Applications, Proceedings, 1977. Edited by R. Ansorge and W. Törnig. IX, 163 pages. 1978.

Vol. 681: Séminaire de Théorie du Potentiel Paris, No. 3, Directeurs: M. Brelot, G. Choquet et J. Deny. Rédacteurs: F. Hirsch et G. Mokobodzki. VII, 294 pages. 1978.

Vol. 682: G. D. James, The Representation Theory of the Symmetric Groups. V, 156 pages. 1978.

Vol. 684: E. E. Rosinger, Distributions and Nonlinear Partial Differential Equations. XI, 146 pages. 1978.

Vol. 690: W. J. J. Rey, Robust Statistical Methods. VI, 128 pages. 1978.

Vol. 691: G. Viennot, Algèbres de Lie Libres et Monoïdes Libres. III, 124 pages. 1978.

Vol. 693: Hilbert Space Operators, Proceedings, 1977. Edited by J. M. Bachar Jr. and D. W. Hadwin. VIII, 184 pages. 1978.

Vol. 696: P. J. Feinsilver, Special Functions, Probability Semigroups, and Hamiltonian Flows. VI, 112 pages. 1978.

Vol. 702: Yuri N. Bibikov, Local Theory of Nonlinear Analytic Ordinary Differential Equations. IX, 147 pages. 1979.

Vol. 704: Computing Methods in Applied Sciences and Engineering, 1977, I. Proceedings, 1977. Edited by R. Glowinski and J. L. Lions. VI, 391 pages. 1979.

Vol. 710: Séminaire Bourbaki vol. 1977/78, Exposés 507–524. IV, 328 pages. 1979.

Vol. 711: Asymptotic Analysis. Edited by F. Verhulst. V, 240 pages. 1979.

Vol. 712: Equations Différentielles et Systèmes de Pfaff dans le Champ Complexe. Edité par R. Gérard et J.-P. Ramis. V, 364 pages. 1979.

Vol. 716: M. A. Scheunert, The Theory of Lie Superalgebras. X, 271 pages. 1979.

Vol. 720: E. Dubinsky, The Structure of Nuclear Fréchet Spaces. V, 187 pages. 1979.

Vol. 724: D. Griffeath, Additive and Cancellative Interacting Particle Systems. V, 108 pages. 1979.

Vol. 725: Algèbres d'Opérateurs. Proceedings, 1978. Edité par P. de la Harpe. VII, 309 pages. 1979.

Vol. 726: Y.-C. Wong, Schwartz Spaces, Nuclear Spaces and Tensor Products. VI, 418 pages. 1979.

Vol. 727: Y. Saito, Spectral Representations for Schrödinger Operators With Long-Range Potentisls. V, 149 pages. 1979.

Vol. 728: Non-Commutative Harmonic Analysis. Proceedings, 1978. Edited by J. Carmona and M. Vergne. V, 244 pages. 1979.

Vol. 729: Ergodic Theory. Proceedings 1978. Edited by M. Denker and K. Jacobs. XII, 209 pages. 1979.

Vol. 730: Functional Differential Equations and Approximation of Fixed Points. Proceedings, 1978. Edited by H.-O. Peitgen and H.-O. Walther. XV, 503 pages. 1979.