STANDARD COSMOLOGY

G. F. R. ELLIS

Department of Applied Mathematics, University of Cape Town
Rondebosch 7700, Cape Town, South Africa

and

School of Mathematical sciences, Queen Mary College
Mile End Road, London E1 4NS, U.K.

1. THE STANDARD MODEL: BASIC FEATURES

The observations underpinning modern cosmology are of two kinds: firstly, observations of individual sources (galaxies, radio sources, qso's, etc) and secondly, measurements of background radiation (radio, microwave, x-ray, etc). As a result of these observations (see e.g. Weinberg's book, 1972, for a summary) we have the following overall picture of large-scale phenomena (Longair, 1985): "Galaxies are the basic building blocks of the universe. Most large-scale high energy phenomena in the universe are associated with the nuclei of galaxies containing a massive object, possibly a black hole. The Big Bang model of the universe is the most convincing framework within which to conduct cosmological research". It is this model with which we are concerned in these lectures.

The major features of the model are, (1) its symmetry, (2) its size (the universe is very big !), (3) there is an evolution of the universe, driven by its present expansion, and (4) the universe had its origin at a beginning a finite time ago. Because of (5) observational problems and (6) causal limitations, certain features are quite unknown, particularly the amount of matter in the universe and its future fate. An unresolved issue is (7) the basic issue of the relation of local physics to the universe and its evolution. Further, (8) fundamental puzzles remain, particularly why features (1)–(4) above should be true.

These lectures concentrate on the main realisation of this model, discussing relevant theory and observations, and problems that remain. We assume standard physics except where otherwise specified; in particular, we assume space-time is 4-dimensional and gravity is correctly described at a classical level by the general theory of *relativity* ("GR"). The general approach of Ellis (1971a) is followed; alternative approaches may be found in Peebles (1971), Weinberg (1972).

1.1 Basic Components

The basic components of any general relativity space-time are (Schutz 1985, Hawking and Ellis 1973) (a) a 4-dimensional *manifold* M, described by coordinates $\{x^i\}$ which can be chosen arbitrarily; (b) a symmetric *metric tensor* $g_{ab}(x^i)$ determining the light-cone structure and the behaviour of clocks in the space-time; (c) a torsion-free *connection*, described by the connection components $\Gamma^a{}_{bc}$ and related to the metric tensor by the condition $g_{ab;c} = 0$, which determines covariant derivatives and so the space-time geodesics and the curvature tensor $R^a{}_{bcd}$; (d) a conserved matter *stress tensor* T_{ab} which is the sum of components $T_{M\,ab}$ (labelled by the index M running from 1 to N) describing the energy, momentum, and stresses of all matter and fields in the space-time, and determining the space-time metric through Einstein's *field equations*

$$G_{ab} \equiv R_{ab} - (1/2)\, R\, g_{ab} = \kappa\, T_{ab} - \Lambda\, g_{ab} \tag{1}$$

where κ is a constant, $R_{ab} \equiv R^c{}_{acb}$ is the Ricci tensor, and $R \equiv R^a{}_a = g^{ab} R_{ab}$ the Ricci scalar, and for generality we have included the "Cosmological constant" Λ (possibly zero). The behaviour of each matter component $T_{M\,ab}$ ($M = 1$ to N) will be determined by *equations of state* describing the physics of that kind of matter by relating its components in a suitable way.

In addition, in a cosmological model there is defined at each space-time point a unique *4-velocity vector* u^a (Ellis 1971a) which is a unit vector:

$$u^a u_a = -1 \tag{2}$$

representing the average motion of matter at that point when a suitable averaging scale is utilised. The world-lines defined by this vector field are the world lines of *fundamental observers*, that is, observers moving with the average motion of matter at each space-time point; u^a is tangent to these world-lines, and if τ is proper time measured along them then $u^a = dx^a/d\tau$. Any particular galaxy will have a motion that is close to, but not identical to, that of a fundamental observer. As a consequence of the existence of this preferred 4–velocity, there also exists at each point a uniquely defined *symmetric 3–metric*,

$$h_{ab} \equiv g_{ab} + u_a u_b, \tag{3}$$

which is a projection tensor ($h^a{}_b h^b{}_c = h^a{}_c$) projecting into the 3–dimensional space orthogonal to u^a ($h^a{}_a = 3$, $h^a{}_b u^b = 0$); thus h_{ab} is just the metric of the rest–space of observers moving with 4–velocity u^a (if X^a and Y^b are both orthogonal to u^a, then $X.Y = X^a g_{ab} Y^b = X^a h_{ab} Y^b$). By (3), the space–time interval ds^2 can be written

$$ds^2 \equiv g_{ij} dx^i dx^j = h_{ij} dx^i dx^j - (u_i dx^i)^2 \tag{4}$$

showing how the displacement (dx^i) is split into a spacelike part (orthogonal to u^i) determined by h_{ij}, and a timelike part (parallel to u^i).

We will now examine the nature of these structural components in a standard universe model, examining features (1) – (8) above in turn.

2. SYMMETRY AND GEOMETRY

The basic feature of the geometry of the standard model is that it is *isotropic about every point*, and consequently is spatially homogeneous. The intuitive concept of isotropy about a point p is that all directions about p are completely equivalent: no observations can distinguish any one direction about p as being different from any other. This can be extended to the idea of *isotropy about a world–line* , i.e. at every moment in an observer's history, his observations are all isotropic. It then is highly plausible that if a space–time is isotropic about more than two world–lines, it must be spatially homogeneous; and indeed this can be rigorously proved. Instead of giving this proof, we follow a standard argument: observations (discussed further below) indicate that on a large scale, the universe is isotropic about us. Then either (1) we are preferred observers, and the universe is observed to be anisotropic about most world–lines; or (2) we are *typical* observers and the universe is isotropic about each world–line. The first possibility is regarded as implausible, so we adopt the second (a variant of what is usually referred to as *the Cosmological Principle*, Bondi 1960, Weinberg 1972);

this leads to the standard model, as we now demonstrate in detail.

2.1 Metric Time Dependence

Consider first the covariant derivative $u_{a;b}$ of $u_a \equiv g_{ab}u^b$. This defines a spatial vector $u^{\bullet a} \equiv u_{a;b}u^b$, the acceleration vector, which is orthogonal to u^a ($u^{\bullet a}u_a = 0$ by (2)); this must vanish at every point or it would define a unique spacelike direction, breaking isotropy, so $u^{\bullet a} = 0$. Since (1) implies $u^a u_{a;b} = 0$, we see that $u_{a;b}$ is completely orthogonal to u^a. In order that it be isotropic about u^a, it is then necessary that $u_{a;b}$ be proportional to the metric of spaces orthogonal to u^a, i.e. $u_{a;b} = \lambda\ h_{ab}$ for some function λ. Taking the trace of this equation we find

$$u_{a;b} = (1/3)\,\Theta\, h_{ab} = u_{b;a}, \quad \Theta \equiv u^a{}_{;a} \tag{5}$$

(which shows that the fluid acceleration, shear σ_{ab} and vorticity ω_{ab} vanish). This implies in particular that $u_{[a;b]} = u_{[a,b]} = 0$ where as usual square brackets denote anti-symmetrization; hence there exists a function $t(x^i)$ such that

$$u_a = -\, t_{,a} \quad => \quad \{t_{,a}h^a{}_b = 0,\ \ t^\bullet \equiv t_{,a}u^a = 1\} \tag{6}$$

the minus sign being introduced for convenience and the last equality following from (2). The function t, determined up to a constant, is a cosmic time function for the universe measuring proper time along every fundamental world line (as follows from (6)). Further if X^a is orthogonal to u^a, then (6) shows $t_{,a}X^a = 0$, i.e. X^a lies in a surface {t = constant}, so these surfaces are just the surfaces orthogonal to the flow lines generated by u^a, i.e. are the surfaces of instantaneity for fundamental observers. Thus t is a synchronised time for fundamental observers; it is convenient to choose it as the time-coordinate x^0. Spatial coordinates x^μ (μ = 1,2,3) are most conveniently chosen as comoving coordinates, i.e. as coordinates that are constant along the fundamental flow lines: $x^{\bullet\mu} \equiv x^\mu{}_{,a}u^a = 0$ (note that x^μ is just a function, so the covariant derivative is the same as the partial derivative). Then the coordinates $\{x^i\}$ are normalized comoving coordinates, with x^μ labelling the fundamental flow lines and t measuring proper time along them, and the 4-

velocity vector u^a has components

$$u^a = \delta^a{}_0 = (1, 0, 0, 0). \tag{7}$$

Putting (6) into (4) then shows

$$ds^2 = h_{ij}dx^i dx^j - dt^2. \tag{8}$$

where $h_{ab}u^b = 0 \Rightarrow h_{a0} = 0$ $(a = 0,1,2,3)$. Note that here and in the sequel, we regard tensors in the 3–spaces {t = constant} as 4–tensors that project into these 3–spaces, e.g. h_{ab} has indices a, b running from 0 to 3 and is such that $(h_{ab})_\perp = h_{ab}$ where the "perpendicular" subscript indicates projection orthogonal; to u^a on all indices. In comoving coordinates this is equivalent to $h_{0a} = 0 = h_{a0}$ $(a = 0,1,2,3)$.

All physical or geometrical functions must be functions of t alone (or else they would not be seen as isotropic by some fundamental observer). In particular, the spatial gradients of Θ must vanish everywhere, so $\Theta_{,a}h^a{}_b = 0$ which implies $\Theta = \Theta(t)$. Now from (7), (8) $u_a = -\delta_a{}^0$ so $u_{a;b} = \Gamma^0{}_{ab} = (1/2)\partial g_{ab}/\partial t$. Thus (5) shows $(1/2)\partial h_{ab}/\partial t = (1/3)\Theta(t)h_{ab}$. Integrating,

$$h_{ab}(x^i) = S^2(t)\, f_{ab}(x^\mu), \qquad S(t) \equiv A \exp \int^t (1/3)\, \Theta(t')\, dt' \tag{9a}$$

The function S(t) is arbitrary by the multiplicative constant A; from (9a), it is related to $\Theta(t)$ by

$$\Theta(t) = 3\, \dot{S}(t)/S(t) \tag{9b}$$

where a dot represents time differentiation: i.e. for any $f(x^i)$, $\dot{f} \equiv f_{,a}u^a = \partial f/\partial t$.

2.1.1 The scale function and expansion

Equations (9) show that S(t) is the scale function for the universe: all spatial distances (measured in a surface t = constant) between pairs of fundamental observers scale with S(t) as t changes, because these distances are determined by the spatial metric (9a) in which time enters only through the common scale factor S(t). In more detail, let γ_1 be any curve in a surface $\{\Sigma_1: t = t_1\}$ joining the world–lines C_1, C_2 of two fundamental observers, and given in terms of the normalised comoving coordinates by $x^a(v) =$

$(t_1, \lambda^\mu(v))$. The distance d_1 measured between C_1, C_2 along this curve will be

$$d_1 = \int S(t_1)\,\{f_{\alpha\beta}(x^\mu)\, d\lambda^\alpha/dv\; d\lambda^\beta/dv\}^{1/2}\, dv. \tag{10a}$$

At any later time t_2 the distance between the same world lines along the corresponding curve γ_2 given by the same functions $\lambda^\mu(v)$, i.e. $\{t = t_2,\ x^\mu = \lambda^\mu(v)\}$ will be given by the corresponding expression with t_1 replaced by t_2 and so will be related to d_1 by

$$d_2 = (S(t_2)/S(t_1))\, d_1 \tag{10b}$$

showing how S(t) is a common scaling factor for spatial distances. The tangent vector η to each such curve γ is orthogonal to u^a and has components $\eta^a = dx^a/dv = (0,\ d\lambda^\mu(v))$ and (because the curves γ_1, γ_2 are dragged into each other by the fluid flow) commutes with $u^a = dx^a/dt = \delta^a{}_0$, so

$$(\eta^\bullet)^a \equiv \eta^a{}_{;b}u^b = u^a{}_{;b}\eta^b = (S^\bullet(t)/S(t))h^a{}_b\eta^b = (S^\bullet(t)/S(t))\ \eta^a \tag{11a}$$

the later equalities following from equations (5) and (9) and the fact that η is orthogonal to u^a. Integrating,

$$\eta^a = S(t)\, K^a, \quad K^{\bullet a} \equiv K^a{}_{;b}u^b = 0, \ \ K_a u^a = 0. \tag{11b}$$

The vector $\varepsilon\eta$ (ε any small constant) is a relative position vector for neighbouring particles along the curves γ. If we split this vector into a direction e^a ($e_a\, e^a = 1$, $e_a u^a = 0$) and a distance δl by the equation $\varepsilon\eta^a = e^a \delta l$, then it follows from equations (11) that

$$(e^a)^\bullet = 0, \tag{12a}$$

$$(\delta l)^\bullet = H(t)\ \delta l, \qquad H(t) \equiv S^\bullet(t)/S(t). \tag{12b}$$

Thus the directions of neighbouring fundamental particles are unchanged (relative to physically non-rotating axes, represented here by a parallelly propagated basis) as the universe evolves, while the "Hubble law" (12b) shows that at any time t_1, the rate of change of distance is proportional to distance, with the same constant of proportionality $H(t_1)$ for all directions. That is, at any time t_1 the expansion is an isotropic expansion determined by the "Hubble constant" $H(t_1)$.

Overall, the conclusion is that the space sections {t = constant} in the standard model are conformally mapped into each other as the universe evolves. In fact equation (5) shows that the vector field u^a is a conformal Killing vector; it generates this conformal mapping between the space sections.

2.2 Metric Spatial Dependence.

To determine the nature of the components f_{ab} we need to note that the 3–spaces {t = constant} with metric h_{ab} given by (9) must themselves be isotropic about each point (or their structure would define spatial directions that would violate the demand of isotropy and would be observable through their effects on geodesics). Hence the only possibility for the Ricci tensor ${}^3R_{ab}$ of these 3–spaces is that it must be isotropic:

$$ {}^3R_{ab} \equiv {}^3R^c{}_{acb} = (1/3)\, {}^3R\, h_{ab} \tag{13a} $$

where ${}^3R$ is the Ricci scalar of these 3–spaces: ${}^3R = {}^3R_{ab}g^{ab} = {}^3R_{ab}h^{ab}$. Now in all 3–dimensional spaces the number of independent components of the Riemann tensor is the same as the number of independent components of the Ricci tensor, so these tensors are completely algebraically equivalent to each other. The equation giving the Riemann tensor from the Ricci tensor is

$$ {}^3R_{abcd} = {}^3R_{ac}h_{bd} - {}^3R_{ad}h_{bc} - h_{ad}{}^3R_{bc} + h_{ac}{}^3R_{bd} - (1/2){}^3R(h_{ac}h_{bd} - h_{ad}h_{bc}) \tag{13b} $$

(this form is required because the right hand side must have the full Riemann tensor symmetries and the correct contractions to give the Ricci tensor and scalar). Now substituting (13a) into (13b) shows

$$ {}^3R_{abcd} = K^*(h_{ac}h_{bd} - h_{ad}h_{bc}), \quad K^* = {}^3R/6 \tag{14a} $$

showing that the 3–spaces with metric h_{ab} are spaces of constant curvature K^*. To determine the functional form of ${}^3R(x^i)$, substitute (13a) into the contracted Banchi identities ${}^3R_a{}^b{}_{;b} = (1/2){}^3R_{;a}$. The component of this equation orthogonal to u^a shows that ${}^3R = {}^3R(t)$; the component parallel then shows ${}^3R(t) = B/S^2(t)$, B a constant. Thus

$$K^* = k/S^2(t) \tag{14b}$$

where k is a constant, the curvature of the 3–spaces with metric $f_{ab}(x^\mu)$ (this identification follows because when S(t) = 1, then $f_{ab} = h_{ab}$ by (9a) and $K^* = k$ by (14b)). Now S(t) can be rescaled by a constant (i.e. we can choose the constant A in (9a) arbitrarily). By equations (9a), (14b), this has the effect of rescaling $f_{ab}(x^\mu)$ and k by constants while leaving unchanged the quantities h_{ab} and K^*, which are invariants (defined uniquely by the space–time geometry) and so are unaffected by this choice. By this rescaling, the constant k can be scaled to +1 if it is positive or −1 if it is negative. This choice will be understood from now on; that is, the 3–space metric $h_{ab}(x^\mu)$ will be a space of constant curvature K^* given by (14b), where k is +1, 0, or −1. In the cases k = +1 and k = −1, S(t) is then uniquely determined, but in the case k = 0, a freedom to rescale S(t) by a constant remains.

2.2.1 Spatial coordinates and geodesics.

The properties of the 3–spaces {t = constant} are determined by equations (14), but for detailed calculations we need to choose explicit coordinates which will determine a specific form for the metric components. Choose any point p in the surface $\{\Sigma_1: t = \text{constant} = t_1\}$ and draw the radial geodesics γ through p in Σ_1 with curve parameter chosen to be the radial distance r as measured by the metric $f_{ab}(x^\mu)$. The actual distance (as measured by h_{ab}) will then be $S(t_1)r$ so (as $S(t_1)$ is constant along each of these curves) r is an affine parameter on each of them. Isotropy about every world line implies the 3–metrics are spherically symmetric about p so the surface $\{S_d: r = d/R(t_1)$ in $\Sigma_1\}$ is a 2–sphere orthogonal to the geodesics γ, with metric proportional to that of a unit 2–sphere. Putting this together,

$$d\sigma^2 \equiv h_{ij}(x^k)\, dx^i\, dx^j = S^2(t)\{dr^2 + f^2(r)\,(d\theta^2 + \sin^2\theta\, d\phi^2)\} \tag{15}$$

where the function of proportionality f(r) is independent of θ, ϕ because of isotropy, and must obey the limit $f(r) \sim r$ as $r \to 0$ because the origin of coordinates is a regular space–time point. To determine f(r) we use the geodesic deviation equation (Synge and Schild, 1961)

$$\delta^2 U^a/\delta r^2 + {}^3R^a{}_{bcd} V^b U^c V^d = 0 \tag{16}$$

for the radial geodesics with tangent vector $V^a = dx^a/dr = \delta^a{}_1$ and connecting vector $U^a = dx^a/d\theta = \delta^a{}_2$. Substituting into (16) from (14) and noting that $V_a U^a = 0$, $V^a h_{ab} V^b = S^2(t)$ shows

$$\delta^2 U^a/\delta r^2 + k\, U^a = 0 \tag{17a}$$

To turn the covariant derivatives into ordinary derivatives, choose a parallel propagated orthonormal basis $\{e_\alpha\}$ along the geodesics where

$$e_1{}^a = S^{-1}(t)\delta^a{}_1, \quad e_2{}^a = \{S(t)f(r)\}^{-1}\delta^a{}_2, \quad e_3{}^a = \{S(t)f(r)\sin\theta\}^{-1}\delta^a{}_3,$$

then the components of the deviation vector are $U^a = S(t)f(r)\delta^a{}_2$ and (17a) becomes

$$d^2f(r)/dr^2 + k\, f(r) = 0. \tag{17b}$$

The solutions with the required limit behaviour at the origin are

$$f(r) = \sin r \text{ if } k = 1,\ \ f(r) = r \text{ if } k = 0,\ f(r) = \sinh r \text{ if } k = -1. \tag{18a}$$

Thus finally the FLRW metric in these coordinates is

$$ds^2 = g_{ij}dx^i dx^j = -\,dt^2 + S^2(t)\{dr^2 + f^2(r)\,(d\theta^2 + \sin^2\theta\, d\phi^2)\} \tag{18b}$$

with f(r) given by (18a).

The nature of these spaces of constant curvature follows immediately from this derivation. From (15), the sphere S_d at distance $d = S(t_1)r$ from the origin of coordinates in Σ_1 has surface area $A_1 = 4\pi S^2(t_1)f^2(r)$. The nature of the space is characterised by the curves of the area $A_1/S^2(t_1)$ against distance $d/S(t_1)$, where we divide out the scaling factor to obtain curves applicable at all times (and representing exactly the geometry of the 3–spaces with metric f_{ab}, obtained when $S(t_1) = 1$). Thus one can imagine testing the geometry of the space sections (in principle) by comparing the radii and surface areas of spheres centred on the origin of coordinates p. In the flat space (k = 0) case, the Euclidean relation holds: the area is proportional to the square of the distance. In the hyperbolic case ($k = -1$), the area increases faster with distance than in the Euclidean case; and in the elliptic case ($k = +1$), the area increases slower than in the Euclidean case.

Two fundamental properties follow from the discussion.

2.2.2 Spatial homogeneity.

Firstly, the point p was an arbitrary point in the surface Σ_1; we could equally well have chosen any other point p' as the origin of coordinates, and (because k in (14b) is a constant) would have obtained the identical geodesic behaviour and metric components centred on that point. Thus the spatial sections (with metric (15)) are completely homogeneous: all points are completely equivalent to each other. From (5) and (9) the scale factor and expansion are also constant on Σ_1 (which is an arbitrary surface of constant time), so the space–time itself (with metric (18)) is spatially homogeneous; and by the argument preceding equation (9) all scalars (e.g. the density and pressure) must also depend only on the time coordinate t. Thus all physical and geometrical properties are identical at all points of each surface {t = constant}, which are therefore surfaces of homogeneity. Hence we have shown that "isotropy everywhere" implies spatial homogeneity of the space–time.

The property of homogeneity can be mathematically formalised in various ways (Ellis and Matravers, 1985); most commonly this is done in terms of the existence of continuous symmetries of the space–time and associated Killing vector fields (see e.g. Weinberg 1972). We have not used this approach to derive the metric, but the result is of course the same. Given the metric (18), one can prove that, for every point p in Σ_1, there exist 3 independent non–zero Killing vector fields which vanish at p and generate the isotropy group around p; and 3 independent Killing vector fields that are non–zero at p, and generate the group of translations of p. Together these Killing vector fields generate the full 6–dimensional group of isometries of a general FLRW space–time, guaranteeing spatial homogeneity and isotropy (there exist extra isometries in the special cases where the space–time itself is a space of constant curvature; these are the cases of de Sitter, Minkowski, and anti–de Sitter spacetimes respectively, only possible if the matter is degenerate as discussed later).

2.2.3 Topology

The discussion so far has related only to local properties; but it should be realised that different global connectivities are possible in each case. In the case $k = 0$, the spatial sections are flat and we can change (r, θ, ϕ) to Euclidean coordinates (x,y,z) in the standard way; the metric will then be

$$ds^2 = -dt^2 + S^2(t)\{dx^2 + dy^2 + dz^2\} \tag{19}$$

Itis usual to assume these coordinates have the standard range: $-\infty < x,y,z < \infty$; then the spatial sections $\{t = \text{constant}\}$ are infinite, without boundary, and of infinite volume; and there is an infinite amount of matter in the universe. However there are many other possibilities. The simplest is the torus topology, where there are numbers L_x, L_y, L_z such that if the point p has coordinates (x,y,z) it is identified with every point q with coordinates $(x + nL_x, y + mL_y, z + pL_z)$ where (n, m, p) are arbitrary integers. In this case each space section $\{t = \text{constant}\}$ is without boundary but is of finite volume, and there is a finite amount of matter in the universe, which has spatially closed ("compact") spatial sections. There are many other possible topologies (see e.g. Ellis 1971b). Thus as well as giving the space-time metric, we need to specify its topology in order to give a full description of its geometry.

The case $k = -1$ is similar: the "natural" topology of the space-sections is that of Euclidean 3-space, but there are many other possible topologies allowing finite, closed spatial sections. When $k = +1$, things are fundamentally different. To see this, consider the geodesics and coordinates described above, where now $f(r) = \sin r$. It is clear that the area of the sphere S_d reaches a maximum when $r = d/S(t_1) = \pi$ and thereafter decreases to zero as $r \rightarrow 2\pi$ at a point q "antipodal" to p. To see what is happening, consider geodesics γ_1, γ_2 leaving p in opposite directions. They intersect each sphere S_d in points r_1, r_2 respectively that are antipodal to each other in S_d; therefore as $r \rightarrow 2\pi$, these geodesics approach q from precisely opposite directions. Hence if one moves from p along the geodesic γ_1, after approaching q and passing through it one continues along the path of the geodesic γ_2 and then arrives back at p; and this happens whatever direction is chosen for γ_1. Thus when $k = +1$ the spatial sections are *necessarily* closed and of finite volume (the situation is exactly modelled by the 2-dimensional surface of an ordinary sphere, which is the 2–dimensional analogue of the 3–dimensional space of constant curvature we are discussing). There are still various topologies possible, but all of them are compact (Ellis 1971b). This is why Einstein preferred this case to the other possibilities: it solves various problems in physics, e.g. what are the boundary conditions on physical fields at infinity ? (this problem falls away when there is no infinity, and periodic boundary conditions are imposed by the topology).

2.3 Relation to Observations

These notes have dwelt at some length on the symmetries and geometry of the space–time metric, because they are the foundation for the other properties of these space–times. A major issue, then, is the question: is the metric (18) in fact a reasonable representation of the universe around us ?

Observational relations will be discussed in some detail later on, but for the present the point is: is the universe in fact spatially homogeneous and isotropic ? On a small scale it is clearly neither. Thus what is implied is some estimate of an averaging scale such that the universe is spatially homogeneous and isotropic on that scale and above. This scale cannot be smaller than about 100 Mpc for we have seen structures in the distribution of matter up to that kind of size (e.g. L'Apparent et al, 1986). A direct observational proof of spatial homogeneity on a large scale is unfortunately very difficult even if we assume that the Einstein field equations hold (Ellis 1980). An indirect approach via the idea of showing all observed matter has had uniform thermal histories (Bonnor and Ellis, 1986) is promising, but runs into theoretical as well as observational problems. Even a direct check of "small–scale" spatial homogeneity, using a definition that can be disproved (cf Stoeger, Ellis, and Hellaby 1987), is difficult to carry out.

As a consequence, the deduction of spatial homogeneity is usually made on the basis of the observed isotropy of matter and radiation about us (when averaged on a sufficiently large scale), and particularly because of the high degree of isotropy of the microwave background radiation. The deduction of spatial homogeneity then takes place on the basis of the assumption that we are not near the centre of the universe, providing the base for the argument for homogeneity sketched in the preceding sections; however one should be aware that this is an unverifiable philosophical assumption (Ellis 1975) and there are alternative possibilities (Ellis 1979), e.g. it could be that the universe is inhomogeneous and we <u>are</u> near its centre ! (see e.g. Ellis et al 1978). Confidence in the resulting standard models is then strengthened by their overall success in relating disparate phenomena (e.g. element abundances and the microwave background radiation temperature, and ages of stellar clusters and the Hubble constant).

While this approach provides a reasonable foundation for believing in an isotropic and spatially homogeneous world model, it should be developed further to give a proper view of the relation between that smoothed out model and the inhomogeneous distribution of matter in the real universe. What is required is some kind of "fitting" procedure which will state how to best approximate a lumpy universe by an exactly smooth model (Ellis and Stoeger 1987a). This requires a detailed analysis of possible observations in both space-times (cf section 5).

2.4 The origin of uniformity

If we accept the standard model, we are led to one of the major problems in cosmology: namely, *why is the universe so smooth ?* The basic issue here is that the standard FLRW models – exactly spatially homogeneous and isotropic universe models – are *a priori* infinitely improbable. Thus one of the major themes of cosmology in recent times is a series of attempts to explain, e.g. through the "inflationary universe" proposal or Penrose' arrow of time argument, how this extraordinarily unlikely situation could have come about. This is in contrast to cosmology of the 1930's to 1970's when it was simply accepted that the universe is uniform because it was created uniformly (to some extent this was regarded as "explained" through the Cosmological Principle).

3. THE SIZE OF THE UNIVERSE

One of the major features of the observed universe is that it is very large. Now of course in a $k = 0$ or $k = -1$ universe with its natural topology, the spatial sections are *infinite*: so it hardly makes sense to talk about the size of the universe. However we can only observe a small section of the universe, and the scale of the part we can see can be estimated from the present value of the Hubble constant (which by (9b) is just one third of the present value $\Theta(t_0)$ of the expansion Θ). The value of the Hubble constant is a matter of debate (Rowan Robinson 1985), but even allowing for this uncertainty the size of this observed region of the universe is of the order of 10^{10} light years. There are two important implications: this figure is very large in relation to man, and it is very, very large in relation to microphysics.

3.1 Observational Restrictions

The scale of the observable region of the universe is very large in relation to man. The total history of astronomy – say 2000 years – is insignificant compared with the timescale of 10^{10} years given by the inverse Hubble time; and if we were now to set of in a spacecraft at the speed of light and travel for 2000 years, we would not have moved very far in our own galaxy, let alone the local cluster of galaxies. Thus, as pointed out by Schucking, Hoyle, and Sachs, on a cosmological scale ***we can only view the universe from one space-time point***, the event "here and now"; we cannot move significantly, either in space or in time, from this event, and we can only see events on its past light cone. This is a fundamental limitation on observational cosmology and our ability to determine the structure of the universe observationally (cf Ellis 1975, 1980).

3.2 The origin of extent

The observable region of the universe is extremely large compared with the natural time and length scales of microphysics, namely the Planck units of 5×10^{-44} secs and 10^{-33} cm (see e.g. Barrow and Tipler 1986). The problem then is, why is the universe so large ? or equivalently, why is it so old ? (Dicke and Peebles 1979). The underlying issue is the relation of the structure of the observed universe to microphysics. If the structure of the universe is essentially determined by microphysical processes, it is difficult to explain its large size relative to these fundamental scales. If initial conditions unrelated to microphysics are responsible, what are those conditions and how do they come to be independent of the fundamental physical laws of nature ? Either way we have a problem; the "Inflationary Universe" idea (Guth 1981), discussed later, proposes a possible solution.

4: THE EVOLUTION OF THE MATTER IN THE UNIVERSE

4.1 Conservation Equations

The evolution of the universe is governed by the gravitational equations (1) provided that appropriate equations of state describe the physics of the matter components. The total matter stress tensor T_{ab} must satisfy the conservation equations

$$T^{ab}{}_{;b} = 0 \tag{20}$$

which are consistency conditions for equations (1); each individual matter component $T_{M\,ab}$ (M = 1 to N) will also individually be conserved provided there is no significant interaction between that matter component and the others. The question then is what are the different significant matter components and what are the appropriate equations of state to use. We have to describe matter (galaxies and possible intergalactic matter) and cosmic background radiation.

4.2 Equations of state

Various quite different matter descriptions are possible.

4.2.1 The kinetic theory approach

This is interesting both in its own right, and as a basis for the fluid approximation. If the distribution of particles is isotropic everywhere in an expanding universe, then in both the Liouville (Ehlers Geren and Sachs 1968) and Boltzmann (Treciokas and Ellis 1971) cases the space–time must be a Robertson–Walker space–time; this provides further justification for using the observed isotropy of the microwave background radiation as evidence for the standard FLRW geometry (Hawking and Ellis 1973). However the converse is not true: there exist exact solutions of the Einstein–Liouville equations in a FLRW space–time in which the distribution function of particles is anisotropic (Ellis Matravers and Treciokas 1983). We will not develop the topic further here; a review of kinetic theory and cosmology will be found in Sachs and Ehlers (1971).

4.2.2 The fluid approximation

A continuum approximation seems sensible in describing the matter in the observed region of the universe, which contains about 10^{11} galaxies each containing of the order of 10^{11} stars. The stress tensor of a fluid can always be written in the form

$$T_{ab} = \mu u_a u_b + q_a u_b + u_a q_b + p\, h_{ab} + \pi_{ab} \tag{21}$$

where $q_a u^a = 0$, $\pi_{ab} u^b = 0$, $\pi^a{}_a = 0$. However in the case of the standard model, the requirement of isotropy demands that both q_a and π_{ab} are zero; then the stress tensor takes the "perfect fluid" form

$$T_{ab} = \mu u_a u_b + p\, h_{ab}, \quad \mu = \mu(t), \quad p = p(t). \tag{22}$$

One should note here the paradox that this *does not necessarily mean* that the fluid obeys the perfect fluid equations of state. For example a simple "imperfect" fluid might obey the Eckart equations of state (Ellis 1971a)

$$q_a = -k\, h_a{}^b(T_{,b} + T\, \dot{u}_b), \quad \pi_{ab} = -\lambda\sigma_{ab} \tag{23}$$

where T is the temperature, $k \geqq 0$ the heat conduction coefficient and $\lambda \geqq 0$ the *viscosity coefficient*. In the case of a FLRW universe, not only will $\dot{u}_b$ and σ_{ab} vanish (as discussed above) but isotropy will also show that $T = T(t)$, so by (23) both q_a and π_{ab} will vanish also, and the stress tensor (21) will take the form (22). A similar conclusion holds for the more realistic Israel–Stewart equations of state (Israel and Stewart, 1979; Hiscock and Lindblom, 1987): in a FLRW space–time, they are consistent with the perfect fluid form (22). It is even possible that a "perfect fluid" stress tensor only makes sense in conditions of zero shear, because in general anisotropic expansion will *generate anisotropic pressures* (Treciokas and Ellis, 1971).

Given the stress–tensor form (22), the physics of the situation is still indeterminate until the relation between p and μ *is specified*, either directly as an equation $p = p(\mu)$, or indirectly in terms of some other appropriate variables (e.g. a temperature T *and entropy* S). The usual approximation, valid for most time in the evolution of the universe, is the "gamma–law"

$$p = (\gamma - 1)\mu, \quad 1 \leqq \gamma \leqq 2 \tag{24a}$$

which includes the case of "pressure–free matter":

$$\gamma = 1 \quad \Leftrightarrow \quad p = 0 \tag{24b}$$

probably an adequate description when the matter component dominates at late times in the universe, and "pure radiation":

$$\gamma = 4/3 \iff p = \mu/3, \quad \mu = aT^4 \tag{24c}$$

probably an adequately description when radiation dominates at early times in the universe. There are times when more detailed descriptions will be required, specifically when significant interactions are taking place between the various matter and radiation components present, or when bulk viscosity is significant; but for most of the time, equations (24) suffice for cosmological modelling purposes.

Because of (5), the only non-trivial conservation equations (20) take the form

$$\mu^\bullet + (\mu + p)\, 3\, S^\bullet/S = 0 \tag{25}$$

in the standard model when the stress-tensor necessarily takes the perfect fluid form (23). With the simple equation of state (24a) this shows

$$\mu(t) = M\,(S_0/S(t))^{3\gamma}, \quad M = \text{constant} \tag{26a}$$

where S_0 is the present value of $S(t)$, leading in particular to the "pressure-free matter" form

$$\mu_m(t) = M_m\,(S_0/S(t))^3, \quad M_m = \text{constant} \tag{26b}$$

and the "radiation" form

$$\mu_r(t) = M_r\,(S_0/S(t))^4, \quad M_r = \text{constant}, \tag{26c}$$

which implies

$$T(t) = T_0\,(S_0/S(t)), \quad T_0 = \text{constant}. \tag{26d}$$

We can normally regard the matter in the universe as comprising a non-interacting mixture of pressure-free matter and of radiation, that is, the total stress tensor is (22) with

$$\mu = \mu_m + \mu_r, \quad p = \mu_r/3 \tag{27}$$

where μ_m, μ_r are given by (26b), (26c); however at times of strong interaction these relations will have to be modified to take into account energy transfer between the different components. Current best estimates suggest that the value of the present matter energy density M_m lies in the range 10^{-31} gm/cc ≦ M_m ≦ 10^{-29} gm/cc, while the present microwave background radiation temperature T_0 = 2.75 K implies that the present radiation energy density M_r has the value $M_r = 10^{-33}$ gm/cc (other known radiation components are negligible relative to *this*). Any of the currently proposed "dark matter" components should obey either (24b) (if non-relativistic) or (24c) (if relativistic), and so can be regarded as an addition either to (26b) or (26c). If the standard values quoted here are correct, the universe is matter dominated ($\mu_m > \mu_r$) now and at late times, but radiation dominated at early times ($\mu_m < \mu_r$ when $S(t)/S_0 < 100$ in the low density case and when $S(t)/S_0 < 10000$ in the high density case).

4.2.3 A scalar field

It is possible that a scalar field Φ of mass m (possibly zero) and with potential $V(\Phi)$ is important in the early universe. If so its equation of motion is

$$\Phi^{;a}{}_{;a} - m^2 \Phi = -\partial V/\partial\Phi \tag{28a}$$

and its stress tensor given by

$$4\pi T_{\phi\alpha\beta} = \Phi_{,a}\Phi_{,b} - (1/2)\, g_{ab}\, (\Phi_{,s}\Phi^{,s} + 2V(\Phi) + m^2\Phi^2)\,; \tag{28b}$$

the conservation equations for this stress tensor are satisfied as a result of the equations of motion (28a) (Lightman et al, 1975). If the field is spatially homogeneous, $\Phi = \Phi(t)$, then this stress-tensor takes the perfect fluid form (22) with

$$8\pi\mu_\phi = \dot{\Phi}^2 + 2V(\Phi) + m^2\Phi^2,\ \ 8\pi p_\phi = \dot{\Phi}^2 - 2V(\Phi) - m^2\Phi^2 \tag{29a}$$

while in a FLRW universe the equation of motion (28a) becomes

$$\partial/\partial t\, (S^3\partial\Phi/\partial t) = S^3\{m^2\Phi - \partial V/\partial\Phi\}. \tag{29b}$$

4.3 Energy and Causality Conditions

Whatever the detailed equation of state chosen for matter, there are two kinds of restrictions that must normally apply. The first is ***causality***: no signal (a sound wave, elastic wave, etc) must be able to move faster than light. This gives the upper limit in (24a); if γ were larger than 2, a sound wave could travel faster than light, violating the basic principles of Special Relativity.

The second kind of restriction is inequalities on the energy density and pressure, collectively known as *energy conditions* (Hawking and Ellis, 1973). For our purposes there are two important such restrictions. Firstly, we normally require that the inertial mass density of matter be positive; in terms of the fluid variables (21), this is the requirement

$$\mu + p \geqq 0. \tag{30a}$$

If this is not true, the matter has very unusual properties, for example from (25a) it will follow that on compressing the fluid its energy density decreases. This implies a mechanical instability. Secondly, we normally require also that the active gravitational mass density be positive; in terms of the fluid variables this is the requirement

$$\mu + 3p \geqq 0. \tag{30b}$$

If this is not true, the gravitational effect of the matter will be negative (as discussed below).

While both inequalities will be satisfied by "normal" matter, they may not be true at various times in the cosmological context. For example, suppose a scalar field were to dominate at early times; then

$$4\pi(\mu_\phi + p_\phi) = \Phi^{\bullet 2}, \quad 4\pi(\mu_\phi + 3p_\phi) = 2\Phi^{\bullet 2} - 2V(\Phi) - m^2\Phi^2 \tag{31a}$$

showing that the active gravitational mass can become negative. As a specific example, if Φ were to remain constant for some time (as is possible if the potential V is appropriately chosen) then

$$\Phi^{\bullet} = 0 \quad \Rightarrow \quad \mu_\phi + p_\phi = 0, \quad \mu_\phi + 3p_\phi = -2\mu_\phi \leqq 0. \tag{31b}$$

This is what happens in some versions of the "inflationary universe" idea: the inertial mass density takes the limiting value of (30a) that (by (25a)) allows expansion but no change of density; and (30b) is violated. One should also note that in this case the equations of motion are indeterminate, and the fields present determine no unique timelike direction. When the strict inequality holds in (30a), the 4-velocity u^a is unique (and consequently the spatial geometry (15) is uniquely determined; particularly one cannot have different choices of k in (18a) by making different choices of u^a).

In general we may expect energy violations when quantum fields are dominant. However problems could arise even in the classical regime, both because of the question of effective averaging of the field equations (Ellis 1984b) and because effective negative pressures could arise in a gravitating fluid, where all inter–particle forces are attractive rather than repulsive.

4.4 The Thermal History of the Universe

The thermal evolution of the universe takes place within the general context set by the relations above. If we follow conditions back into the past, as S(t) decreases towards zero, by (26d) the radiation temperature increases without limit. Three major physical processes then take place.

Firstly, the average photon energy eventually exceeds the binding energy of any bound system; so successively only simpler and simpler structures can exist. In particular, above about 1500 K molecules cannot exist, any that form being immediately broken down into their component atoms. Above about 3000 K, atoms can no longer exist, as their ionisation energy is exceeded, and they are broken down into nuclei and electrons, creating a plasma. Above about 10^8 K, complex nuclei cannot exist, being broken down into neutrons and protons. At high enough energies, the baryons themselves may be decomposed in to their constituent quarks.

Secondly, a series of interactions which exchange energy between particles but cannot take place at low temperatures become possible, and proceed fast enough to create thermal equilibrium between different components of the primeval plasma. In particular above about 3000 K Thomson scattering of photons and electrons equalizes their temperatures, forcing the matter to follow the radiation temperature law, because the thermal capacity of the radiation vastly exceeds that of the matter at least until the matter becomes relativistic, when it follows the same law as radiation. This scattering makes the universe opaque to photons at early times. Similarly above about 10^9 K electron–neutrino scattering brings neutrinos into thermal equilibrium with the other components, making the universe opaque to neutrinos.

Thirdly, a series of interactions transmuting particles into each other take place at high temperatures and create equilibrium abundances between particles species; for example, above about 6×10^9 K neutrons and protons are transmuted into each other by weak interactions and attain an equilibrium abundance. At high enough temperatures, pair production thresholds are exceeded so that as the temperature rises, photon pairs successively create equilibrium abundances of electron–positron pairs; muon pairs and pion pairs;

proton and anti-proton pairs; and so on. At each such threshold, radiative energy is converted to energy of relativistic particles.

A fourth process will take place if our present understanding of the nature of fundamental forces is correct: as the temperatures increases, successively forces that are disparate at lower temperatures will become unified at higher temperatures. Thus the electromagnetic and weak forces will be unified above the electroweak unification temperature; the strong force will also be unified with them above the grand unification temperature; and gravitation too will become unfied with them above the Planck temperature.

Together these processes create equilibrium abundances of particles at very early times, the nature of the equilibrium being determined simply by the conserved quantities at that temperature; at temperatures below the GUT temperature but above 10^{10} K, these are electric charge; baryon number; and lepton number.

In the standard model, the temperature drops from indefinitely high values at the origin of the universe to the present value of 3 K, and as it does so the processes mentioned above, dissociating complex structures and establishing equilibrium when the temperatures are high enough, one after the other cease to be effective. Also particle pair annihilation takes place, converting the energy of relativistic particles to radiative energy, and forces which were unified at higher temperatures separate into the fundamental forces we experience today as the temperature drops below the various unification thresholds.

Two important consequences follow. Firstly, the universe (which was opaque) becomes transparent to different forms of radiation as the temperature drops – notably to neutrinos at about 6×10^9 K, and to photons at about 3000 K (the time of *decoupling* of matter and radiation). Secondly, as the temperature drops the mechanisms maintaining thermal equilibrium successively cease to function and the expansion of the universe then allows non-equilibrium processes to take place, building up complex structures. In particular baryo-synthesis will take place at about the Grand Unified Energy (if our current understanding of Grand Unified Theories is correct) and nucleosynthesis will take place at about 10^9 K, creating the light elements (deuterium, helium, and lithium); heavy elements are formed later in the interiors of massive stars. We will not discuss these complex physical processes further here, referring the reader to Weinberg (1972), Turner and Schramm (1979), and Schramm (1983) for details.

4.5 Observations

The details of this evolution depend on the rate of expansion of the early universe, determined by the Einstein field equations (discussed in the next section). The evidence for the picture outlined is truly impressive.

Firstly, the 3K cosmic background radiation with a black body spectrum is evidence for a hot early stage of the universe, this radiation being interpreted as relic radiation of that era of thermal equilibrium. Two histories are possible. In the first, this radiation decoupled from matter at a time t_d when the temperature was about 3000 K and (by (26d)) the scale factor $S_d \equiv S(t_d)$ was $S_d = 10^{-3}S_0$, and the radiation has travelled freely ever since. In the second, decoupling took place as above but a dense intergalactic medium then heated up to about 10^6 K at fairly recent times, causing significant Thomson scattering again and resulting in our seeing back only to an era when $S = 10^{-1}S_0$.

Secondly, element abundances are predicted by the theory of nucleosynthesis in the early universe and can be compared with current observations of element abundances. Theory predicts production of He^4 mass fractions of between 0.1 and 0.3 in the early universe (depending on the number of neutrino types and the present baryon density) while the values estimated from present day observations lie between 0.20 and 0.29, in good agreement with the theory; indeed the observations can be used to restrict the number of neutrino types (at most one neutrino-like particle is at present unknown) and to limit the present baryon density to relatively low values ($M_{baryon} \leq 10^{-30}$ gm/cc).

4.6 Why does the matter in the universe evolve ?

A dominant feature of the present-day universe is that it is both presently out of thermal equilibrium, and is full of the products of a lack of equilibrium in the past (light and heavy elements; molecules; radioactive substances; stars; organic and biological systems). One of the major questions facing cosmology is how this can be.

The standard model solves the issue in effect by referring it back to the evolution of the universe: local systems can be out of equilibrium and evolve because the universe itself is evolving (Marx and Sato 1986). This leaves the question of why the universe itself should be evolving; which we return to after studying Einstein's equations.

Alternative models based on a non-evolving ("Steady-State") universe have to develop some approach that tackles the evolution issue satisfactorily. The original Steady-State theory (see e.g. Bondi 1960) does so through a continuing (steady-state) operation of non-equilibrium processes (in that case, the creation of matter); static inhomogeneous universes may also be able to do so in a similar way (Ellis Maartens and Nel 1978). However some proposals for stationary universe (e.g. Segal's theory, 1976) seem unable to explain the fundamental feature of the continuing existence of non-equilibrium processes and the evolution of the contents of universe. A theory that cannot do so is unable to describe important aspects of the observed universe.

5: THE ORIGIN OF THE UNIVERSE

The evolution of the universe is governed by the Einstein field equations (1). We consider in this section how one of these equations implies that the standard universe model must evolve from a singular origin; and in the following section, the complete set of equations governing its evolution.

5.1 Instability and Evolution

The uniquely defined 4-velocity vector field u^a must satisfy the Ricci identity

$$u^a{}_{;bc} - u^a{}_{;cb} = u^d R_{dabc} \tag{32}$$

Substituting into this relation from equations (5), (9b) and (6), we find

$$u^d R_{dabc} = S^{\bullet\bullet}(t)/S(t)\,(h_{ac}u_b - h_{ab}u_c) \tag{33a}$$

Using the Ricci tensor definition $R_{db} = R_d{}^a{}_{ba}$ this shows

$$u^d R_{db} = 3\,S^{\bullet\bullet}(t)/S(t)\,u_b, \quad u^d u^b R_{dabc} = -\,S^{\bullet\bullet}(t)/S(t)\,h_{ac} \tag{33b}$$

Equations (33b) show that $u^d u^b R_{db} = -\,3\,S^{\bullet\bullet}(t)/S(t)$. Using equations (1) and (22) to evaluate $u^d u^b R_{db}$ in terms of μ, p and Λ, we find

$$3S^{\bullet\bullet}(t)/S(t) + \kappa/2\,(\mu + 3p) - \Lambda = 0 \tag{34}$$

which is the ***Raychaudhuri equation*** for the standard models. This is the basis of the instability proof and the singularity theorem.

5.1.1 Instability of the static solution

A static solution will correspond to S(t) = constant, so $S^{\bullet\bullet} = 0$. Then (34) shows

$$\kappa/2\ (\mu + 3p) = \Lambda, \tag{34a}$$

which shows that (provided the energy condition (30b) holds) *a necesary condition for a non-empty static solution is* $\Lambda > 0$ (in fact this was the reason Einstein introduced the cosmological constant into the field equations, for in 1917 everyone knew that the universe was static). Given an equation of state satisfying the energy condition (30a), (30b) and a positive value of Λ, there will be some value S_{*} of S(t) such that (34a) is satisfied. If S(t) is then perturbed to a slightly larger value, μ will decrease. Thus provided $\partial p/\partial \mu > -1/3$ (which will in particular be true if p is constant); then $\kappa(\mu + 3p) < 2\Lambda$ so by (34a) $S^{\bullet\bullet} > 0$, and the solution will expand to even larger values. Similarly if it is perturbed to slightly smaller values it will continue decreasing. Thus, as discovered by Eddington in 1930, *a static FLRW solution is unstable*. This discovery was the reason that the idea of an expanding universe became accepted amongst relativists, and it enables us to "explain" the fact that the universe is evolving.

Two comments are in order. Firstly, these results will not hold if the inequality (30a) is not strictly violated (e.g. if equality holds then perturbing the scale function results in no change in density) or if (30b) does not hold (if this sign is reversed, then Λ need not be positive and the solution can be stable). Secondly, the idea of a perturbation of the universe is a logical absurdity: as (by definition) there is no physical system outside the universe, there is no way it can be perturbed. Nevertheles the stability argument has powerful appeal to physicists, and is still accepted as implying that the univers must either expand or contract, and therefore that it must be evolving. This can be interpreted in the sense that the universe is <u>not</u> exactly a FLRW spacetime, and so can be regarded as "perturbed" from a FLRW universe in that it is a little different from those models.

5.1.2 The Singularity Theorem

Observed redshifts show the universe is presently expanding (the present Hubble constant $H(t_0) \equiv H_0$, given by (12b), is positive). Thus if $\Lambda = 0$ and the energy inequality (30b) holds, $S^{\bullet\bullet} < 0$ and so S(t) increases monotonically from zero; in fact then

$$S(t) \to 0 \text{ a finite time } t_0 \text{ ago where } t_0 < 1/H_0. \qquad (*)$$

This shows that the universe not merely evolves, but has a beginning at a singular origin at a "Hot Big Bang" (HBB) where (by (26)) the energy density and temperature diverge. One should note here particularly that high pressures will not prevent the initial singularity in the standard model, where there are no spatial pressure gradients and the fluid moves geodesically. On the contrary, pressure enhances collapse because of its contribution 3p to the active gravitational mass (μ+3p) in equation (34).

In principle a non-zero value of Λ, if it is larger than the active gravitational mass density $\kappa(\mu + 3p)$ in (34) can prevent this origin at an initial singularity . However in practice this will not work if we believe the usual explanation of the microwave backgroudn radiation as relic radiation form a hot big bang. The reason is that Λ must dominate equation (34) at the time of the turn-around. However the radiation is evidence of a thermal equlibrium when the temperature was at least 3000 K, i.e. when $S/S_0 < 1/1000$; but at that time, by (26) $\mu > 10^9 \mu_0$. If Λ were large enough to overcome this, it would be the dominant force in the solar system today and would certainly have been detected.

One possible way out of the prediction of an initial singularity is to suppose that the energy conditions (30) are violated at early enough times when quantum field effects come into play (cf section 4.3). Another is that the Einstein field equations might be incorrect, and some alternative field equations might hold at very high densities leading to an effective energy condition violation then (e.g. when quantum gravity effects are dominant). Both are possible ways of avoiding the initial singularity in the universe (cf Ellis 1984b for a summary and references).

5.1.3 Ages

The singularity statement (*) already indicates the possibility of an ***age problem***: if the age of any object (stars, rocks, elements, or whatever) in the universe is larger than $1/H_0$ then that component of the universe is older than the universe itself, which cannot be. Again the possible ways out are a violation of the energy conditions; alternative field equations; or a positive cosmological constant. The difference from singularity avoidance is that to avoid singularities, the effective energy violation must occur at early times (when $S/S_0 < 1/1000$), while to have a serious effect on ages the effective energy violation must occur at late times. Sharper age inequalities than (*) can be derived by more detailed examination of the field equations (cf the following section).

5.2 The problem of the origin of the universe

The prediction (*) raises the problem, why is there a beginning to the universe ? In the 1930's cosmologists such as Eddington and Lemaitre at first found it very difficult to believe this could be so. The point is that what is predicted is not merely creation of matter in a given spacetime, but creation of space and time, and even of the laws of physics themselves. Space-time curvature diverges at this event, so this is a *space-time singularity* (cf Tipler et al, 1980). Furthermore if $k=-1$ or $k=0$ with the natural topology, what is implied is the creation of an infinite amount of matter at that time. How can we account for this creation of all from nothing ?

A partial response is to consider the alternatives: suppose the universe were indeed a steady-state universe that existed forever. Is this really a preferable explanation of the "origin" of the universe: that it existed forever ? If one considers this, it seems just as problematic as the coming into being of the universe a finite time ago. Aesthetic and philosophic issues are at stake: physics cannot decide. In any case it is clear that non-quantum energy violations will not avoid the prediction of an initial singularity where the classical equations break down; there must be a period in the early history of the universe where quantum gravity effects dominate. These may or may not succeed in explaining the origin of the universe from some previous structure in a non-singular way, but at least from a classical viewpoint the universe has a singular origin.

One of the issues highlighted by this discussion is the difficulty – actually impossibility – of testing the field equations, or the equations of state of matter, in the very extreme conditions relevant to the very early universe. Thus it is worth studying the nature of the standard universe models without imposing particular field equations ("cosmography") as well as with particular gravitational equations imposed ("cosmology"). While one can provide plausibility arguments for particular field equations, *proving* their validity is not possible; so it is difficult, in the end, to disprove any behaviour that is allowed cosmographically.

6: THE EVOLUTION OF THE UNIVERSE

6.1 Equations

We have considered so far only one of the field equations; but all of them must hold. To ensure all equations are satisfied we now need to check the ten field equations and four conservation equations.

6.1.1 The conservation equations

The energy conservation equation is (25); these determine the evolution of the energy density μ once suitable equations of state are specified (cf section 4.2). The momentum conservation equations are identically satisfied because there are no pressure gradients to cause non-geodesic motion, and the fluid flow lines are indeed geodesics.

6.1.2 The (0,0) field equation

The (0,0) field equation is (34). This is a second order equation. Now using (25a) one can show that $(S^2\mu)^\bullet = -SS^\bullet(\mu + 3p)$. Thus multiplying (34) by $S^\bullet$ and integrating we find the first integral

$$3S^{\bullet 2} - \kappa\mu S^2 - \Lambda S^2 = \text{const} \equiv E \tag{35}$$

valid whenever $S^\bullet \neq 0$.

6.1.3 The $(0,\mu)$ field equations

From (33b) (1) and (21) we find $u^d R_{db} h^b{}_c = 0 = q_c$ showing that the three $(0,\mu)$ field equations are identically satisfied for FLRW universes with a perfect fluid matter source (22) (which is required by isotropy).

6.1.4 The (ν,μ) field equations

The six (ν,μ) equations determine the geometry of the 3-spaces {t = constant} orthogonal to the fluid flow in the following way. The metric ${}^3g_{ab}$ of these 3-spaces is the total projection into them of the 4-dimensional metric, i.e. ${}^3g_{ab} = h_a{}^c h_b{}^d g_{cd} = h_{ab}$. Similarly the covariant derivative ${}^3\nabla^a$ in the 3-spaces is defined as the total projection of the 4-dimensional covariant derivative $\nabla_a \equiv {}_{;a}$; e.g. ${}^3\nabla_c T_{ab} \equiv h_c{}^f h_a{}^d h_b{}^e \nabla_f T_{de}$. It follows from this definition that ${}^3\nabla_a$ is indeed the correct Christoffel connection for the 3-space metric h_{ab}, i.e.

it is torsion-free and preserves that metric: ${}^3\nabla_c h_{ab} = 0$.

The curvature tensor ${}^3R_{abcd}$ is defined in the standard way by the 3-dimensional covariant derivative; i.e. for every vector field X^a in the 3-surfaces {t = const} (<=> $X_a u^a = 0$),

$${}^3\nabla_c {}^3\nabla_b X_a - {}^3\nabla_b {}^3\nabla_c X_a = X_d \, {}^3R^d{}_{abc}. \tag{36a}$$

To work this out, it is important to remember that each 3-derivative is the total projection of the corresponding 4-derivative, and so

$${}^3\nabla_c {}^3\nabla_b X_a = h_c{}^f h_b{}^d h_a{}^t \, \nabla_f \, (h_d{}^e h_t{}^r \nabla_e X_r). \tag{36b}$$

Using (3) and (5), equations (36) lead to the Gauss-Codacci equations for a FLRW universe:

$${}^3R_{abcd} = (R_{abcd})_\perp \; + \; (1/9)\, \Theta^2\, (h_{ad}h_{bc} - h_{ac}h_{bd}). \tag{37}$$

Contracting with h^{bd}, the first term on the right becomes $(h^{bd}R_{abcd})_\perp$ = $(R_{ac})_\perp + R_{abcd}u^b u^d$. The former is found from (1) and (22) and the latter from (33c). On using (34) we rederive equations (13a) where now

$${}^3R = 2\,\{\kappa\mu + \Lambda - (1/3)\Theta^2\}. \tag{38}$$

Thus (13a) are 6 of the Einstein field equations provided (38) determines ${}^3R$. Comparing (38) with the first integral (35) shows these two equations are identical provided that ${}^3R = -2E/S^2$; then (14) shows the constant of integration E is related to the constant k of (14) by E = −3k, so (36a) is

$$3S^{\bullet 2}/S^2 - \kappa\mu - \Lambda \;=\; -\,3k/S^2 \tag{38a}$$

where k = +1, 0 or −1; this is the ***Friedmann equation*** for the standard model. We now see that this equation is both a first integral for the Raychaudhuri equation (in consequence of the energy conservation equation) and is also the Gauss-Codacci equation (38) giving the Ricci scalar ${}^3R$ in terms of the energy density μ. Equations (13) and (14) will be valid if the metric takes the form (15). Thus all the (μ,ν) field equations will be valid if the metric is (15) and equation (38) holds where ${}^3R = 6k/S^2$, i.e. if (38a) is valid.

6.1.5 The complete set of equations

The Einstein equations guarantee that if the universe is initially a FLRW model, then it will remain a FLRW universe later on (cf Hawking and Ellis 1973). From the above it follows that all the non-trivial field equations and conservation equations for the standard model will be satisfied if the metric is given by (15) and additionally,

(A) In the exceptional case of a *static universe*, $\Theta = 0$ and μ is constant; (34a), (38a) with $S^{\bullet} = 0$, and suitable equations of state hold;

(B) In the general case of a *non-static universe*, $S^{\bullet} \neq 0$; (25a), (38a), and suitable equations of state hold ((34) will automatically be satisfied as a consequence of the other equations).

6.2 The Einstein static universe

By the above, a static solution to the field equations satisfies (34a) and

$$\kappa\mu + \Lambda = 3k/S^2 \tag{39a}$$

Substituting (34a) into (39a) shows

$$k/S^2 = 3\kappa/2\ (\mu+p) > 0 \tag{39b}$$

giving the general form of the *Einstein static universe* with positive cosmological constant (if (30b) holds) and positive spatial curvature (k = +1 if the (30a) is satisfied). Einstein particularly approved the finite spatial sections of this universe because it solves the problem of boundary conditions at infinity for local physics (Einstein 1956). We reject this as a universe *model both because it does not expand (so there are no redshifts)* and it is unstable (section 5.1.1).

6.3 Initial conditions

In the expanding universe models, it is convenient to parametrise solutions by present-day values of the Hubble constant H_0 and (dimensionless) deceleration parameter q_0, defined by

$$H_0 \equiv (S^{\bullet}/S)_0, \quad q_0 \equiv -\,(1/H_0^2)\,(S^{\bullet\bullet}/S)_0 \tag{40a}$$

together with the effective density parameters

$$\Omega_m = (\kappa\mu_m)_0/3H_0^2,\ \ \Omega_r = (\kappa\mu_r)_0/3H_0^2,\ \ \Omega_\Lambda = \Lambda/3H_0^2 \tag{40b}$$

for the matter and radiation (see equations (26)) and cosmological constant Λ respectively. In terms of these quantities the photon to baryon ratio h is given by $h = \Omega_r/2\omega\Omega_m$ where $\omega = 3.7m_b/2kT_0$; here m_b is the baryon mass, k the

Boltzmann constant, and T_0 the present background radiation temperature. The present value of Ω_m lies between 1 and 0.04 while the present value of Ω_r is about 10^{-4} if the 3K black body radiation is the dominant component (as is suggested by present observations), hence $\mathcal{h} = 10^8$ to 10^{10}.

Evaluating (34) at the present time and using these definitions shows

$$q_0 = 1/2\,(\Omega_m + 2\Omega_r) - \Omega_\Lambda \tag{41a}$$

while (38a) similarly shows

$$(K^*)_0 \equiv k/S_0^2 = H_0^2\,(\Omega_m + \Omega_r + \Omega_\Lambda - 1). \tag{41b}$$

Equation (41a) can in principle be used to estimate Λ and (41b) to estimate $(K^*)_0$ from observations (cf section 9.11 in Rindler 1977). If we assume $\Lambda = 0$ (which is compatible with the observations), the standard assumption of dominant non–relativistic matter ($\Omega_r << \Omega_m$) leads to the relations

$$q_0 = 1/2\,\Omega_m, \quad (K^*)_0 \equiv k/S_0^2 = H_0^2\,(\Omega_m - 1); \tag{42a}$$

showing the critical density separating $k = +1$ and $k = -1$ universes is $\Omega_m = 1$ corresponding to $q_0 = 1/2$. However the assumption of a universe presently dominated by relativistic particles ($\Omega_r >> \Omega_m$) leads to

$$q_0 = \Omega_r, \quad (K^*)_0 \equiv k/S_0^2 = H_0^2\,(\Omega_r - 1). \tag{42b}$$

showing the critical density is again $\Omega_r = 1$ but now corresponding to $q_0 = 1$. These relations determine the value of S_0 if $k \neq 0$; if $k = 0$, that value is indeterminate (S can be rescaled by an arbitrary constant, cf section 2.2)

For later use, it is convenient to rewrite the Friedmann equation (38a) in terms of these quantities and the normalised scale variable $y \equiv S(t)/S_0$. The general result for a matter plus radiation universe is

$$y^{\cdot 2} = H_0^2\,\{\Omega_m/y + \Omega_r/y^2 + \Omega_\Lambda y^2 - k/S_0^2 H_0^2\}. \tag{43a}$$

In the matter dominated case ($\Omega_r = 0 = \Omega_\Lambda$) this reduces to

$$y^{\cdot 2} = H_0^2\,\{2q_0/y - (2q_0 - 1)\} \tag{43b}$$

while in the radiation dominated case ($\Omega_m = 0 = \Omega_\Lambda$) it reduces to

$$y^{\bullet 2} = H_o^2 \{q_o/y^2 - (q_o-1)\}. \tag{43c}$$

One application is that the age t_0 of the universe can be found from the equation $t = \int dy/y^\bullet$, where the integral is taken from $y = 0$ to $y = 1$.

6.4 $\Lambda = 0$ Solutions

6.4.1 Qualitative behaviour

Following the universe back into the past, if the energy conditions are satisfied it has a singular origin at an initial singularity (see section 5.1.2) provided the space–time can be extended back that far, and indeed this is possible (Collins and Ellis 1979, section 4). If $k = -1$ or $k = 0$, the universe expands forever in the future because by (38a) $S^\bullet$ is never zero; furthermore the scale function $S(t)$ is unbounded. If $k = +1$ then for certain equations of state violating (30b) the universe might expand asymptotically to a finite radius, or there could even be a singularity in $S^{\bullet\bullet}$ (Barrow et al 1986); however if $p \geqq 0$, $\mu \geqq p$ it will reach a maximum value of S where $S^\bullet = 0$ and then recollapse to a future singularity where the density and temperature again increase without limit and space–time comes to an end (at least on a classical view). In particular this will happen for a pure "dust" (pressure–free matter) universe, or pure radiation universe. It is noteworthy that the question of whether the universe recollapes in the future or not turns out to be identical as to whether it has spherical space sections ($k = +1$) or not ($k = 0$ or -1), which in turn corresponds to it exceeding the critical density ($\Omega_m + \Omega_r > 1$) or not. However note that while an "open" universe (with infinite space sections) must not exceed the critical density ($\Omega_m + \Omega_r \leq 1$) and will expand forever, a "closed" universe can also expand forever, the simplest example being a $k = 0$ universe with torus topology (section 2.2.3) and exactly critical density.

6.4.2 Exact solutions:

The general behaviour just discussed holds for all reasonable matter. To obtain exact solutions we need to consider particular equations of state. A ***non-interacting mixture of matter and radiation*** is represented by equations (26), (27) leading to (42b); when $\Omega_\Lambda = 0$, this can be written as

$$y^{\bullet 2} = 1/(y^2 S_0^2)\,\{\beta^2 + 2\alpha y - k y^2\} \tag{44a}$$

where

$$\alpha \equiv S_0^2 H_0^2 \Omega_m/2, \quad \beta \equiv (S_0^2 H_0^2 \Omega_r)^{1/2} \tag{44b}$$

The general solution to (44a) may be obtained in terms of the conformal time parameter τ defined through the equation $\tau = 1/S_0 \int dt/S(t)$. The solution in the three cases $k = +1$; $k = 0$; and $k = -1$, is

$$k = +1: \quad y = \alpha\,(1 - \cos\tau) + \beta \sin\tau \tag{45a}$$

$$k = 0: \quad y = \alpha/2\ \tau^2 + \beta\,\tau \tag{45b}$$

$$k = -1: \quad y = \alpha\,(\cosh\tau - 1) + \beta \sinh\tau \tag{45c}$$

where correspondingly (setting $t = \tau = 0$ when $S = 0$)

$$k = +1: \quad t = S_0 \{\ \alpha\,(\tau - \sin\tau) + \beta\,(1 - \cos\tau)\} \tag{4fd}$$

$$k = 0: \quad t = S_0 \{\ \alpha/6\ \tau^3 + \beta/2\ \tau^2\} \tag{45e}$$

$$k = +1: \quad t = S_0 \{\ \alpha\,(\sinh\tau - \tau) + \beta\,(\cosh\tau - 1)\} \tag{45f}$$

It is interesting how in this parametrisation the dust and radiation decouple. Equations (44) of course include as special cases the general pure dust and radiation solutions, corresponding respectively to $\beta = 0$ and $\alpha = 0$.

Particular cases allowing simpler representation are of interest. In the low density case ($k = -1$) there is a mathematically simplest universe model given by the condition that the bracket on the right hand side of equation (44a) is a perfect square. Then $\beta^2 = \alpha^2$, i.e. $\Omega_m^2/\Omega_r = 4/S_0^2 H_0^4$. Using (41b) this becomes

$$\Omega_m = 2(\Omega_r^{1/2} - \Omega_r)$$

giving a value for the present matter density $(\mu_m)_0$ of about 2×10^{-31} gm/cc, very close to the observed value.

The simplest pure matter case is the *Einstein–de Sitter* universe ($\mu_r = 0$, $k = 0$; $q_0 = 1/2$ in (43b)). In this case the scale function is

$$S(t) = A\,(t - t_*)^{2/3} \tag{46a}$$

where A, t_* are constants. Its age is found to be

$$t_0 = 2/3\ \ 1/H_0. \tag{46b}$$

The corresponding simplest pure radiation universe ($\mu_m = 0$, $k = 0$; $q_0 = 1$ in (43c)) has

$$S(t) = A\,(t - t_*)^{1/2}, \qquad (47a)$$

$$t_0 = 1/2\;\; 1/H_0. \qquad (47b)$$

For many purposes, the dynamics of the universe are adequately described by equations (46) at late times when the universe is matter dominated, and by (47) at early times when it is radiation dominated. In general A can be normalised to 1 and t_* to zero without loss of generality, but if one wishes to describe a single model by (46a) at early times and (47a) at lates times, then at the change-over time t_d both $S(t)$ and $S^{\bullet}(t)$ must be continuous (cf Ellis 1987); this is achieved by appropriate choices of the constants A, t_* (e.g. choosing A = 1 and $t_*=0$ in (46a), valid for $t \leqq t_d$, and then finding A, t_* in (47a), valid for $t \geqq t_d$, from continuity of S, $S^{\bullet}$ at t_d). More accurate calculations need the exact equations (44a) or solutions (45).

For completeness, the simplest empty universe model is the ***Milne universe*** ($\mu_r = \mu_m = 0$) given by (43b) or (43c) with $q_0 = 0$. Of necessity, $k = -1$; one finds

$$S(t) = (t - t_*), \qquad (48a)$$

$$t_0 = 1/H_0 \qquad (48b)$$

(the normalisation factor is necessarily A = +1 because k = +1). This is in fact the flat space-time of Special Relativity with a cloud of particles expanding uniformly into it; gravity does not curve the space-time at all (this is possible because it is empty).

While it is possible that the late universe is dominated by relativistic particles, it is more plausible that it is matter dominated, so the solutions of (43b) are of particular significance. The parameters for these models may usefully be taken as the Hubble constant H_0 and the matter density Ω_m (or equivalently $q_0 = \Omega_m/2$). The set of such models and observational constraints on them are discussed by Gott et al (1974, 1976). Direct observational limits from observed helium abundances, ages and densities suggest the universe is a low-density universe: "a variety of arguments strongly suggest that the density of the universe is no more than a tenth of the value required for closure"

(Gott et al 1974). The present observational evidence is the same as at that time; however this conclusion should be reviewed in the light of the possible presence of a large mass of non-baryonic matter that does not affect helium production (Schramm 1983), but there is no direct observational evidence for the presence of such matter.

Exact solutions describing other kinds of matter can be obtained; for example solutions with bulk viscosity are given by Treciokas and Ellis (1971); and kinetic solutions by Ehlers, Geren and Sachs (1968) and Ellis, Matravers and Treciokas (1983). We will not discuss these kinds of solutions further here, although they may through *interesting light on the possible* behaviour of FLRW universe models.

6.4.3 Approximate solutions

If the universe expands to arbitrarily large values of S at late times, $k = 0$ or $k = -1$. The asymptotic solution depends on this value. If $k = -1$, the (curvature) k-term will dominate the Friedman equation (38a) at late enough times for normal matter, cf. (44a) for large y. Thus the asymptotic form of the equation is just $y^{\bullet} = 1/S_0$, leading to the Milne solution (48a) as the asymptotic form. If $k = 0$, the matter term (given by α in (44a)) will dominate at late times, leading to the Einstein–de Sitter solution (46a) as the asymptotic form.

At early times,the matter term will dominate in (38a), *cf (44a) for small y.* Thus the effective equation at early times will be

$$3\, S^{\bullet 2} = \kappa \mu S^2. \qquad (49a)$$

Using the equation of state (24a) and corresponding density behaviour (26a), *the solution is*

$$S = A\,(t - t_*)^{(2/3\gamma)}, \quad \kappa\mu = (4/3\gamma^2)\ 1/t^2 \qquad (49b)$$

showing that the initial expansion depends only on γ (there are no free constants in S(t)). For the plausible case of a radiation dominated early universe ($\gamma = 4/3$), we find

$$S = A\,(t - t_*)^{1/2}, \quad \kappa\mu = 3/4\ 1/t^2, \ T = (3/4a)^{1/4}\ 1/t^{1/2} \qquad (49c)$$

showing the unique relation between temperature T and time t in the early universe which leads to the standard nucleosynthesis predictions.

6.5 $\Lambda \neq 0$ Solutions

While the cosmological constant Λ can be regarded as a separate term in the field equation (1), it can alternatively be thought of as an extra contribution $(T_\Lambda)_{ab} = -\Lambda g_{ab}$ to the matter stress tensor T_{ab}. If so, then by (22), (3) it is equivalent to a perfect fluid with

$$\mu_\Lambda = -p_\Lambda = \Lambda/\kappa = \text{constant} \tag{50a}$$

implying

$$\mu_\Lambda + p_\Lambda = 0, \; \mu_\Lambda + 3p_\Lambda = -2\Lambda/\kappa \tag{50b}$$

Thus this "fluid" always obeys the exceptional limit of (30a) which (by (25a)) allows expansion without change of μ; and violates (30b) if $\Lambda > 0$.

6.5.1 Quantitative behaviour

If $\Lambda < 0$, it acts as an extra attractive force tending to slow down the expansion of the universe to the future (energy condition (30b) is satisfied). The cosmological constant will eventually dominate the Friedmann equation and cause a recollapse to a second singularity in the future. The universe starts at an initial singularity and ends at a second one.

If $\Lambda > 0$, it acts as an extra repulsive force, violating the energy condition (30b). The result depends on the value of k. If $k \leqq 0$, $S^\bullet$ cannot be zero; these universe already escape to infinity when $\Lambda = 0$, so when $\Lambda > 0$ they do so more easily. All these universe start at an initial singularity and expand forever.

If $k > 0$, a wide variety of behaviour is possible. (a) The Einstein static universe is now a possibility at some radius S_c given by (34a), depending on the value of Λ and the equation of state of matter. For $S > S_c$, $S^{\bullet\bullet} > 0$ and the curve S(t) bends up; for $S < S_c$, $S^{\bullet\bullet} < 0$ and the curve S(t) bends down. Thus solutions are possible (b) that start asymptotically close to the Einstein static universe in the past, either (i) with $S > S_c$ and which expand away from it to infinity, or (ii) with $S < S_c$ that collapse away from it to a singularity in the future. Solutions can occur (c) decreasing from an infinite value of S in the past, either (i) collapsing to a singularity in the future, (ii) asymptotically approaching S_c from above, or (iii) decreasing to a finite value $S_m > S_c$ and then re-expanding to infinity. The latter universes are known as "oscillating universes"; they are non-singular, as are (a), (b)(i), and (c)(ii). Finally (d) solutions can start at an initial singularity, either (i) recollapsing in the future, (ii) asymptotically approaching the Einstein static universe in the future, or (iii) expanding forever. In the latter case the universe can slow down near the radius S_c and spend a long time there; such universes are known as *Eddington-Lemaitre* universes. It has not been emphasized above, but the equations are time-symmetric and for each solution there is a time-reversed solution as well; in particular (b)(i) is the time reverse of (c)(ii), (c)(i) that of (d)(iii), and (b)(ii) that of (d)(ii). Robertson (1933) gave a detailed analysis of these possibilities; an extension of the classification *to the case of negative pressures violating* the energy conditions is given by Harrison (1967).

At late time the equation of state may be taken as that of "dust" ($p = 0$). A detailed classification of such models in terms of the observational variables (t_0, $\sigma_0 \equiv 1/2\ \Omega_m$) is given by Rindler (1977, section 9.11). An equivalent analysis in terms of the variables (q_0, σ_0) is given by Stabell and Refsdal (1966), who depict the phase plane for the universe models in terms of these variables and give detailed information on the age of the universe in these models. A large positive Λ term can lead to ages much greater than $1/H_0$ (the limit when Λ is zero).

6.5.2 Exact solutions:

Exact solutions can be given in the dust case in terms of elliptic integrals, but they are not particularly illuminating. However simple exact solutions can be found in the case of an empty model ($p = \mu = 0$) driven by a *positive* cosmological constant Λ. Equation (34) becomes $S^{\bullet\bullet} = -(\Lambda/3)\ S$ with solution $S = A \exp(\omega(t-t_*)) + B \exp(-\omega(t-t_*))$ where $\omega = (\Lambda/3)^{1/2}$ and A, B can be rescaled by choice of the constant t_*. Equation (38a) then gives the relation: $4AB\omega^2 = k$. Thus we find on suitable choice of constants:

$$k = +1 \Rightarrow S = (3/2\Lambda) \cosh \omega t \tag{51a}$$

$$k = 0 \Rightarrow S = A \exp \omega t,\ A \text{ constant}, \tag{51b}$$

$$k = -1 \Rightarrow S = (3/2\Lambda) \sinh \omega t \tag{51c}$$

These are all forms of the space–time of constant positive curvature, the *de Sitter* universe, with the coordinates determined by (51a) covering the whole space–time but the others covering only parts of it (Schroedinger 1955). The same space–time can be represented as different FLRW universes because the choice of 4–velocity vector u^a is not unique in this case; the exceptional equation of state (50b) is precisely the one for which the stress tensor defines no unique timelike eigenvector (cf section 4.3).

Solution (51b) is a *Steady State Universe* explicitly exhibiting the stationary nature of this space–time (the metric form is invariant under the rescaling $t \to t' = t + t_0$, $A \to A' = A \exp(-\omega/t_0)$). However unless matter obeys the exceptional equation of state (50b), this can only be a solution of Einstein's equations if it is empty but has a positive cosmological constant (as above). Bondi, Gold and Hoyle proposed it as a cosmological model in which the Einstein's field equations (and conservation equations (25a)) were abandoned, a continuous creation of matter taking place as the universe expanded (cf Bondi 1960). This has turned out to be incompatible with observations of radio–source number counts (and the microwave background radiation is difficult to explain in this case). Furthermore the original argument for this space–time in terms of its very high symmetry (the "Perfect Cosmological Principle") has been diluted by the realisation that it is geodesically incomplete in the past, and indeed a non–scalar curvature singularity occurs at the boundary of the universe a finite distance from every event in the space–time (Ellis and King, 1974).

It has recently been realised that the exceptional equation of state (50b) could result from quantum fields in the early universe, e.g. from a scalar field Φ (cf section 4.3). This has led to the proposal (Guth 1981) of "Inflation", that is, a period in the early universe where (50b) (a "false vacuum") holds and (51b) describes the expansion of the universe for a sufficient time that the scale function S(t) increases by a very large factor (say e^Z where Z >> 100). This is attractive in many ways, particularly in providing a possible physical explanation for the fact that the observed universe is large (section 3.2). A complete description of the expansion of the univese must then consist of an initial radiation dominated phase (47a), followed by an inflationary phase (51b), followed by a radiation dominated phase (47a), followed by a matter dominated phase (46a). Each phase must be matched to the following one by choosing the constants so that S and $S^\bullet$ are both continuous (cf Ellis and Stoeger 1987b). An important effect of the expansion is to make the "curvature" term k/S^2 in (38a) negligible at the present time; thus by (41c), assuming $\Lambda = 0$, we should live in a high density universe at the present time with the density Ω_m very nearly at the critical value 1.

6.5.3 Approximate solutions

The early solutions with $\Lambda \neq 0$ will be the same as when $\Lambda = 0$. The late evolution of any universe with $\Lambda > 0$ that expands forever will be dominated by that term, and it will tend asymptotically to the de Sitter universe (51b).

6.6 Alternative Field Equations:

Einstein's field equations are an approximation that will not hold at very early times when quantum gravity effects are dominant; and indeed they may be inaccurate at quite recent times, or on scales other than the solar–system scale which is the scale on which our experimental tests validate them. There will of course be a set of cosmological solutions corresponding to each gravitational theory. In most cases the gravitational field equations will be the General Relativity equations (1) plus some extra terms on the left hand side; these terms can be transferred to the right hand side so that they are efefctive matter terms which in turn (as in the case of the cosmological constant, cf. equations (50)) can be expressed as effective contributions to the energy density μ and pressure p in the FLRW universe. Thus alternative field equations are in effect equivalent to alternative equations of state.

Many papers have been written on the effect of such alternative theories on cosmology. In particular Jones (1974) has described in detail all FLRW cosmologies based on theories of gravity that lead to autonomous systems of equations for cosmology. However this will not include theories with higher–order Lagrangians. The cosmological implications of gauge theories of gravity are summarised by Goenner and Muller–Hoissen (1984).

6.7 Observations

6.7.1 Element abundances

The early universe timescale (49c) is tested through the theory and observations of element abundances. As mentioned above, these are in good agreement.

6.7.2 Densities and q_0

The relation between Ω and q_0 is given by equations (41). Observations are not good enough to determine q_0 adequately to distinguish between $k = +1$ and $k = -1$, even if we assume $\Lambda = 0$: they only show $-1 < q_0 < 1.5$ (Sandage and Tamman 1986). Observations of the energy density μ are also unable to determine k (Peebles 1986); specifically because of the numerous forms of "dark matter" (ordinary or exotic) that might be present and undetected.

6.7.3 Ages

In a universe model with vanishing cosmological constant, we may suitably work out the ages on the basis of the "dust" equation of state because this is the plausible equation at late times, and (by (34)) most of the age of the universe is accumulated in this matter-dominated era, the proper time elapsing in the early radiation dominated era being very small. Exact expressions for the age in terms of H_0 and q_0 can be obtained from (43b), but simple estimates will suffice for our purposes. In a high density universe ($\Omega = 1$), (46b) will be a good estimate whereas in a low density universe the limit (47b) can be used as an estimate. For any $p = 0$, $\Lambda = 0$ model the age can be expected to lie within these bounds. A very high density universe ($\Omega > 1$) will give even lower estimates for the age, as will a universe dominated by radiation at late times (which can be estimated from (47b)).

The Hubble constant is unknown to a factor of 2, lying in the range 50 Km/sec/Mpc to 100 Km/sec/mpc. This is close to giving significant limits on universe models. For example, Penny and Dickens (1984) estimate the age of the globular cluster NGC 6752 as 16 ± 2 Gyrs. The lower bound of 14 Gyr implies an upper limit of 70 km/sec/Mpc on the Hubble constant in the low-density limit (48b) and of 47 Km/sec/Mpc in the high density case (46b). If one follows Sandage (1982) in adding $0.2\ H_0^{-1}$ to the age of the globular cluster to allow for the formation of galactic nuclei prior to cluster evolution, then in the low density case $H_0 < 56$ km/sec/Mpc whereas in the high density case $H_0 < 30$ km/sec/Mpc, in apparent disagreement with the observations. If the lower bound on ages came down to 16 Gyr the disagreement would be more marked, providing evidence against the critical density (inflationary universe) proposal; and the presence of relativistic particles at late times makes the problem worse (as (47b) then applies). A way out is to assume the cosmological constant is positive but then it becomes a substantive issue as to why Λ should have such a small non-zero value.

It is not clear that there is an age problem, but there is certainly close to being one at least in the high density cases. On the one had one needs better estimates of the Hubble constant, and on the other of the ages of stars, planets, and elements (see e.g. Audouze 1980 for a summary).

6.8 The particular initial conditions

The evolution of the universe raises a series of questions that have been the concern of cosmology since the 1930's. Firstly, is the density of matter and radiation present greater or less than the critical density ? Equivalently, what is the future behaviour of the universe ? What is its spatial curvature, i.e. is $k = +1, 0,$ or -1 ?

Given answers to these questions, the deeper issues remain: why do the initial values of the universe correspond to these particular parameters ? An attack on that problem is the inflationary proposal, which if correct shows that for a wide variety of initial conditions the universe should be very close to the critical density at the present time. However it unfortunately does not make a prediction as to the value of k, and so is agnostic as regards the future evolution of the universe and the nature of spatial curvature. A second daring attack is via quantum cosmology (Hawking 1984), giving an explanation for some of these parameters in terms of quantum processes and in particular predicting that the universe is a high density universe with closed spatial sections. It is to the credit of these theories that they are open to verification, for example through age tests which are close to showing we can only live in a high-density universe if $\Lambda > 0$.

7: NULL-CONE OBSERVATIONS

Cosmological models only attain a relation to reality through being tested by comparison with observations. It is thus important to deduce all possible observational tests of these models. Some have been mentioned already in the previous sections; this section focusses on light-cone observations in cosmology, i.e. observations made by electromagnetic radiation (optical, radio, X-ray, ultra-violet or infra-red) travelling towards us at the speed of light.

7.1 Equations

Electromagnetic radiation travels on null geodesics in space-time, i.e. on curves $x^a(v)$ for which

$$k^a{}_{;b}k^b = 0, \quad k^a k_a \equiv -\varepsilon = 0, \quad k^a \equiv dx^a/dv. \tag{52}$$

These curves can also be derived as solutions of the Euler–Lagrange equations

$$\partial L/\partial x^a - d/dv\, \partial L/\partial x^{*a} = 0 \tag{53a}$$

where

$$L^2 \equiv g_{ab}(x^a)\, x^{*b}\, x^{*b} = -\, t^{*2} + S^2(t)\, \{r^{*2} + f^2(r)\, (\theta^{*2} + \sin^2\theta\, \phi^{*2})\}, \tag{53b}$$

and $x^{*a} \equiv dx^a/dv$, with first integral

$$L = -\varepsilon \;= \text{constant} \tag{53c}$$

the equivalence of (52) and (53) following from the Christoffel relations for the connection components $\Gamma^a{}_{bc}$, and the specific form for L from (18b).

Because of the spatial homogeneity of these universes, we can consider the radiation to travel on radial null geodesic through through the spatial origin of coordinates $r = 0$. Then $\theta = \theta_1 =$ const, $\phi = \phi_1 =$ const. on the geodesics $x^a(v)$ (having started out radially, spherical symmetry ensures that they continue moving radially; this is obvious from the symmetry but also follows from the geodesic equations), i.e. $d\theta = 0 = d\phi$. Also $\varepsilon = 0$ is equivalent to $ds^2 = 0$; so on these geodesics, (18b) reduces to $dt^2 = S^2(t)\, dr^2$. Thus

$$u \equiv r_0 - r_1 \;= \int dt/S(t) = u(t_1,t_0) \tag{54a}$$

along such geodesics from $r = r_1$ to $r = r_0$, where the integral is taken from the initial time t_1 to the final time t_0 (this result also follows more formally from (53)). In terms of the variable $y \equiv S(t)/S_0$, this relation can be written

$$u(y_1,y_0) \;=\; S_0 \int dy/yy^* \tag{54b}$$

where the integral runs from $y_1 \equiv S_1/S_0$ to 1.

7.1.1 Redshift

Let pulses of radiation be emitted by G at times t_1, $t_1' \equiv t_1 + \Delta t_1$, and be received by O at times t_0, $t_0' \equiv t_0 + \Delta t_0$ where source G is comoving with the expanding fluid at fixed r–coordinate value r_1 and observer O comoving with the expanding fluid at fixed r–coordinate value r_0. Then the observed time dilation for all events at G as seen by O is $\Delta t_0/\Delta t_1$. In particular this applies to the period $\Delta t = c\lambda$ of light of wavelength λ, with the observed cosmological redshift z_c defined by $z_c \equiv \Delta\lambda/\lambda$. Equation (54a) applies to both pulses, with a fixed value of u. Equating the integral from t_1 to t_0 to that from t_1' to t_0', we see that for small Δt_1, Δt_0,

$$1 + z_c = \lambda_2/\lambda_1 = \Delta t_0/\Delta t_1 = S(t_0)/S(t_1). \tag{55a}$$

showing that the redshift depends only on the total expansion of the universe between when the light was emitted and when it was received.

It is characteristic of the redshift effect that this frequency shift is independent of the wavelength of the light observed. We must be careful, however. In general neither the source nor the observer will be exactly comoving (and there could even be local gravitational contributions). Let z_G be the local redshift of a source G' as measured by a comoving observer (for whom G, at the same event, is stationary), and z_O the local redshift of light from a comoving object as measured by the observer O' (at the same event as O). As each redshift corresponds to an observed time dilation, the total observed redshift z for G' as measured by O' is then given by

$$1 + z = (1 + z_O)(1 + z_c)(1 + z_G) \tag{55b}$$

(cf Harrison and Noonan 1979). No direct observation by the observer can distinguish these three contributions to the observed redshift; they have to be inferred indirectly. This is one of the substantial problems in cosmology, contributing to the redshift controversy (Field Arp and Bahcall, 1973).

7.1.2 Observed solid angles and angular diameters

Consider light travelling from the extended source G at r = u radially towards the observer O situated at the origin r = 0 of spatial coordinates. For simplicity we consider the light to travel on the bundle of null geodesics bounded by the coordinate values $(\theta, \theta + d\theta)$ and $(\phi, \phi + d\phi)$. Then the observed solid angle is $d\Omega \equiv \sin\theta d\theta d\phi$. If the light is emitted at a time t_1, the corresponding lengths subtended by this bundle of light at the source (evaluated from (18b) with $dr = 0$, $dt = 0$) are $S(t_1)f(u)d\theta$, $S(t_1)f(u)\sin\theta d\phi$ where u is given by (54), so the cross-sectional area is $dA = S^2(t_1)f^2(u)d\Omega$. Thus if we define the ***area distance*** r_0 by the relation

$$dA = r_0^2 \, d\Omega \tag{56a}$$

then we see that

$$r_0^2 = S^2(t_1)\, f^2(u). \tag{56b}$$

Because of the isotropy of these models, the area distance will also show how all angles behave: an object of linear size $l(t_1)$ will be observed to subtend an angle α where

$$l(t_1) = r_0\, \alpha. \tag{56c}$$

To obtain specific formula for the area distance, one must assume a specific equation of state and use the Friedmann equation (42) to determine y^* (this is how the matter in the universe affects the focussing of null geodesics via the Einstein field equations). In particular if $\Lambda = 0$, $p = 0$, one can integrate using (43b) to obtain Mattig's formula

$$r_0 = H_0^{-1} q_0^{-2} (1+z)^{-2} \left\{ q_0 z + (q_0-1)[(1+2q_0 z)^{1/2} - 1] \right\} \tag{57a}$$

when $q_0 \neq 0$, and

$$r_0 = 1/2H_0 \;\; \{1 - (1+z)^{-2}\} \tag{57b}$$

when $q_0 = 0$. These curves, which should be a good description of the behaviour of observed angular diameters in the late universe, are plotted by Sandage (1961). In principle one can find q_0 by comparing these formulae with observations of angular sizes of distant objects with clearly defined linear scales. It is possible the late universe is dominated by relativistic particles or radiation; one can use (43c) to obtain the formula

$$r_0 = H_0^{-1} q_0^{-1} (1+z)^{-2} \left\{ [q_0(1+z)^2 - (q_0-1)]^{1/2} - 1 \right\} \tag{57c}$$

for this case where $p = \mu/3$. It turns out that the curves obtained from this formula are indistinguishable from those obtained from (57a) for small red-shifts, i.e for $z < 2$ (Ellis and Tivon 1985).

A particularly interesting feature of these formulae, first discovered by Hoyle, is that *refocussing* occurs: any pair of null geodesics diverging from the observer back into the past reach a maximum distance apart and then start reconverging (thus the whole past light cone starts reconverging, see Hawking and Ellis 1968). This is evidenced by ***minimum angular diameters***: if we consider moving an object of fixed linear size l back into the past, its image will decrease in apparent size to a minimum angular diameter at some redshift z_*, and will then start increasing in size again. In particular the simplest case is the high density dust Einstein–de Sitter universe where

$$r_0 = 2/H_0\ (1+z)^{-3/2}\ \{(1+z)^{1/2} - 1\}; \tag{57d}$$

(the particular case of (57a) where $q_0 = 1/2$); here refocussing occurs at $z_* = 5/4$. We have seen qso's and even some galaxies beyond this redshift. In low density dust universes refocussing occurs at higher redshifts, e.g. if $q_0 = 0.02$ then $z_* = 3.97$; thus searching for the refocussing redshift z_* is a possible test for the density of matter in the universe (Ellis and Tivon 1985).

7.1.3 Intensity formulae

(a) ***Free propagation***. Consider light spreading out spherically from the source G at time t_1 towards the observer. The argument above shows that at time t_0 the cross-sectional area of light emitted at the source into solid angle $d\Omega$ is $dA = S^2(t_0)f^2(u)d\Omega$. The fraction of the light emitted by the source that is intercepted by this solid angle is $(d\Omega/4\pi)$, so if the source luminosity is $L(t_1)$, the rate of emission of light into this solid angle is $L(t_1)\ (d\Omega/4\pi)$. As the light propagates freely and photons are conserved, this light will arrive at the observer. The rate of reception of the light will be measured by him to be down by a factor $(1+z)$ from the rate of emission, because of the time–dilation (55a); and the energy per photon will also be measured to be decreased by a factor $(1+z)$, because this energy is proportional to the photon frequency which will be redshifted by this amount. Putting this together, the measured flux of radiation from the source, i.e. the energy received per unit time per unit area by the observer, will be

$$F = (L(t_1)/4\pi(1+z)^2)\ (1/S_0^{\,2}f^2(u)) \tag{58a}$$

By (56b) and (55a), this can be re–expressed as

$$F = (L(t_1)/4\pi(1+z)^4)\,(1/r_0^2) \tag{58b}$$

Thus the ***apparent magnitude*** $m \equiv -2.5 \log_{10}F + \text{const}$ measured for the source observed will be

$$m = -2.5 \log_{10}L(t_1) + 5 \log_{10}\{r_0\,(1+z)^2\} + \text{const.} \tag{58c}$$

On substituting specific expressions for r_0 we obtain predicted magnitude–redshift curves for this universe; e.g. (57b) gives the standard curves, plotted by Sandage (1961), for pressure–free matter. These can in principle be used to determine q_0 by measuring the redshifts and apparent magnitudes of distant sources whose intrinsic luminosity L can be well estimated.

This formula refers to the total radiation received from a source at all wavelengths. In practice the detector will only measure light in a limited waveband. Hence to actually compare prediction and observation we need to know the source spectrum $\mathcal{J}(\nu_G)$ specifying the fraction of light emitted by the source at frequency ν_G into the spectral band $(\nu_G, \nu_G + d\nu_G)$. This spectrum will be red–shifted to the observed frequency ν. The received radiation in the frequency band $(\nu, \nu + d\nu)$ will then be

$$F_\nu d\nu = (L(t_1)/4\pi(1+z)^3)\,\mathcal{J}(\nu(1+z))/r_0^2 \tag{59}$$

where a factor $(1+z)$ allows for the redshifting of the emitted band width $d\nu_G$ to the observed bandwidth $d\nu$. It is preferable to use this formula for comparison with observation than applying "corrections" to equations (58).

Furthermore, (59) is only directly applicable to point sources. For an extended source, at each point of the image a detector will directly measure the observed ***specific intensity*** of radiation I_ν, i.e. the specific flux per unit solid angle. By (56a), (59) this is given by

$$I_\nu d\nu \equiv F_\nu d\nu/d\Omega = (I_G/(1+z)^3)\,\mathcal{J}(\nu(1+z))\,d\nu \tag{60}$$

showing that the observed specific intensity of radiation is independent of the area distance r_0, depending only on the spectrum $\mathcal{J}$, observed redshift z, and source surface brightness $I_0 \equiv L/4\pi dA$ where dA is the cross-sectional area of the part of the source observed in the solid angle $d\Omega$.

Equation (60) applies to any freely propagating radiation; in particular, it applies to black body radiation, and it follows directly from this form that if radiation is emitted as black body radiation at a temperature T_e, because its spectrum at the source then takes the form $\mathcal{J}(\nu) = \nu^3 b(\nu/T_e)$ where b is the Planck function for black body radiation at temperature T_e, it will be observed as black-body radiation at a temperature T where

$$T = T_e/(1+z) \tag{61}$$

if z is the redshift of the matter emitting the radiation. By (55a), this agrees with the temperature law (26d) for radiation in the expanding universe.

(b) *Absorbtion and emission.* In general there will be absorbtion and emission of radiation between a discrete source and the observer. A general formula for such processes can be given (see e.g. Narlikar and Davidson 1966, Ellis 1971a, Harrison 1977) which becomes particularly simple for processes such as Thomson scattering (Bahcall and Salpeter 1965, Bahcall and May 1968). In effect, these approaches are based on the Boltsmann equation in the radiative transfer limit. An alternative approach to the effect of scattering processes on background radiation is via a local cell theory approach (Harrison 1977) that ties it more closely to the thermal history of matter and radiation.

7.1.4 Number Counts

If the observer O at time t_0 detects dN sources lying between r-coordinate distances u and u+du in a solid angle $d\Omega$ (the observed light being emitted by the sources at about the time t_1), the volume dV in which the sources lie is $dV = dA\ dl$, where the area $dA = r_0^2 d\Omega$ by (56a) and the length $dl = S(t_1)du$ by (18b). Thus if $n(t_1)$ is the number density of sources and p the probability that sources in this volume are detected by the observational procedure, then

$$dN = p\ n(t_1)\ dV = p\ n(t_1)\ S(t_1) r_0^2 d\Omega\ du = p\ n(t_1)\ S(t_1)^3 f^2(u) d\Omega\ du$$

that is

$$dN = p\ (n(t_1)/(1+z)^3)\ S(t_0)^3 f^2(u) d\Omega\ du \tag{62a}$$

which can be compared with observation, because u can be found in terms of the observable quantity $r_0(1+z)$ by use of (56b):

$$u = f^{-1}\left\{r_0/S(t_0)(1+z)\right\}, \tag{62b}$$

where f^{-1} is the inverse of the function f defined in (18a), and $S(t_0)$ can be written in terms of H_0, q_0 by use of equations (41). If galaxies are conserved during the time interval (t_0,t_1) then $n(t_1) = n(t_0)(1+z)^3$ and so the total number of galaxies N detected from $u = 0$ to $u = u_1$ is

$$N = n_0 S_0^3 d\Omega \int p\, f^2(u')\, du' \tag{62c}$$

where the integral from 0 to u can be easily performed if the detection probability p is independent of u.

7.1.5 The Observational map

The basic relations given above are the basis of detailed observational tests which depend on statistical analysis of observed galactic images, and so in turn on the detection and selection criteria for these images.

Whether or not a particular source is detected depends on the size and intensity of its image. Suppose the intrinsic surface brightness I_G in (60) varies with distance ρ across the surface of a galaxy as

$$I_G = B_0(z)\, \not{f}(\rho/a(z)), \quad \not{f}(0) = 1 \tag{63}$$

where B_0 is the central brightness, $\not{f}(\zeta)$ the brightness profile and $a(z)$ the scaling radius; examples of suitable profiles are the Hubble profile $\not{f}(\zeta) = (1+\zeta)^{-2}$ and the de Vaucouleurs profile $\not{f}(\zeta) = \exp(\zeta^{1/\kappa})$ where $\kappa = 1$ for spirals and 4 for ellipticals. By (56c), ρ is related to the observed angle β from the centre ($\rho = 0$) of the galaxy image by

$$\rho = r_0\, \beta\,. \tag{64}$$

With (63) and (60) this shows the observed specific intensity will be

$$I_\nu(\beta, z) = \mathcal{B}(z)\, \not{f}(\beta\chi), \tag{65a}$$

where

$$\chi(z) \equiv r_0(z)/a(z), \quad \mathcal{B}(z) \equiv B_0(z)\mathcal{I}(\nu(1+z))/(1+z)^3. \tag{65b}$$

Thus the specific flux of radiation measured through an aperture of semi-angle A centred on the galaxy will be

$$F_\nu(A,z) = \mathcal{B}(z)/\chi^2 \; g(A\chi(z)) \tag{66a}$$

where

$$g(\beta) \equiv \int 2\pi\xi \, f(\xi)\, d\xi \tag{66b}$$

the integral being taken from 0 to β. One can work out the integral explicitly for the profiles mentioned above.

A simple characterisation of detection limits is through a specific intensity limit $S_L(\nu)$: the galaxy image is detected at frequency ν up to where it fades away into the noise as its specific intensity (65a) drops below the value $S_L(\nu)$. Then the *apparent angle* A_{ap} of the image is given by solving the equation

$$S_L(\nu) = I_\nu(A_{ap}, z) \tag{67a}$$

From (65a) this gives

$$A_{ap}(\nu,z) = 1/\beta\chi \; f^{-1}\{S_L(\nu)/\mathcal{B}(z)\}, \tag{67b}$$

where f^{-1} is the inversion of the profile f. The *apparent flux* of the source F_{ap} is then given from (66a) by

$$F_{ap}(\nu,z) = F_\nu(A_{ap}, z) \tag{68}$$

and the apparent magnitude m_{ap} is defined by $m_{ap} = -2.5 \log_{10} A_{ap} + \text{const.}$

This is an isophotal magnitude up to the observed brightness limit $S_L(\nu)$.

The luminosity L, and so absolute magnitude M, of a galaxy with surface brightness distribution (63) is determined directly from the central brightness $B_0(z)$ and scale size $a(z)$. If it is situated at redshift z in a universe with area distance $r_0(z)$, one can determine the apparent size A_{ap} and apparent magnitude m_{ap} of its image from (67), (68) and so determine the

observational map: (M, a) → (m_{ap}, A_{ap}) from the <u>object plane</u> with coordinates (M, a) to the <u>image plane</u> with coordinates (m_{ap}, A_{ap}). This map determines the image characteristics from the source characteristics. This map is very non-linear; many examples are given by Ellis, Perry and Sievers (1984), which also takes into account point-spread effects in the detector. It is suggested that *selection effects* determining the fraction p of galaxies measured by a galaxy counting machine or included in a catalogue, depend on the apparent magnitude and apparent angle, i.e. on the image location in the image plane; further details are given in Ellis, Perry and Sievers (1984).

7.2 Observational data

The relations above, together with analysis of local astrophysical properties of objects observed, form the basis of the detailed observational tests discussed for example in Balian et al (1980), Setti and van Hove (1984), Kolb et al (1986).

(a) Analysis of observations of galaxies and other discrete sources involves detailed statistical analysis of their properties, in particular knowledge of the "luminosity function" giving their luminosity distribution. Unfortunately statistical variation in source properties together with the problem of source evolution, i.e. knowing the intrinsic parameters of the source at the time the observed light was emitted, remains a major stumbling block in using the standard tests: the (magnitude, redshift), (angular diameter, redshift), (number count, magnitude), or (number count, redshift) relations or variants such as the (Volume, luminosity) test. These give useful restrictions (Gott et al 1974, Loh 1986) but are far from definitive in determining q_0 and $(K^*)_0$.

As a consequence, attention has recently been more and more on statistical analyses of source properties, e.g. of the source covariance function, and on dynamical properties that relate to the mass distribution, e.g. galaxy rotation curves, virial theorem mass estimates, local velocities and their relation to mass inhomogeneities, and galaxy formation processes (see e.g Peebles 1986). Local gravitational lensing may also be a way of determining mass densities (see e.g. Subramian and Chitre 1987). These all take place in the context of the expanding universe model and observational properties outlined above, and should take this context properly into account; for example covariance function analyses are sometimes calculated without taking into account cosmological curvature effects, but the extension to do so is possible (Groth and Peebles 1977, Bonometto and Lucchin 1978). These analyses also are begining to take into account the non-smooth distribution of matter in the universe, i.e. the existence of clusters and voids on a very large scale and associated peculiar velocities.

(b) The other main stream of observations is detailed analysis of the background radiation spectrum at all wavelengths and the associated interaction of radiation with matter (see e.g. Rees 1980, Wilkonson 1984). Like dynamical estimates, this helps materially to place limits on the various possible forms of "dark matter" and so to obtain important density estimates, particularly when combined with nucleosynthesis data and theory (see e.g. Schramm 1983, Matzner 1986). However there are so many theoretical possibilities that this data also does not uniquely determine q_0, $(K^*)_0$.

7.3 Observational issues

The prime observational issues are the nature and evolution of luminous matter on the one hand and the nature and density of dark matter on the other. Uncertainties about both kinds of matter prevent us from determining the cosmological parameters directly from observations. In particular, we do not known the matter density in the universe, nor do we known how to estimate source evolution. An alternative would be if we could locate sources whose intrinsic properties were well–determined by local physical processes, and so whose luminosity and size could be estimated directly, thereby providing a calibrated "standard candle". It is possible that supernovae will fulfil this role, but the erratic behaviour of the Magellanic Cloud supernova is not promising in this regard.

As observational knowledge becomes more detailed, we are more and more faced with the problem of how to sensibly fit a completely smooth universe model (as described here) to a lumpy reality. Up to now, this is being done in a rather ad hoc manner. Some initial proposals for a more systematic approach (Ellis and Stoeger 1987a) need development.

A final point is that it is possible we are seeing the same galaxies many times over through multiple images; indeed it is possible we have already seen all the matter in the universe, and that a number of the "distant galaxies" we see are in fact images of our own galaxy. Such a "small universe" is possible in each family of FLRW models, i.e. whether $k = +1$, $k = 0$, or $k = -1$. Their observational properties are intriguing; relatively few galaxies can give an impression of many galaxies approximately uniformly distributed over the sky (Ellis and Schreiber 1986). The underlying issue is that a local determination of the universe model (knowledge of q_0, H_0, etc.) does not determine the topology of the universe in the large (section 2.2.3); and that topology therefore needs observational determination.

8: CAUSAL PROPERTIES

The FLRW universe models are conformally flat, so their local causal properties are the same as that of flat space-time. Specifically, these models are stably causal and so the "future" of each event (the space-time region that can be reached by future directed timelike or null curves from that event) is bounded by the future null cone (Hawking and Ellis 1973, Tipler et al 1980); similarly the past is bounded by the past null cone. As a consequence the region of the universe that can influence or be influenced from any event is strongly limited.

Consider the event R at time $t = t_0$ situated on the world line of the observer O at the origin of spatial coordinates $(r = 0)$. Because causal influences can travel at most at the speed of light, an event S to the past of R at coordinates (t_1, r, θ, ϕ) can causally affect R if and only if $r \leqq u(t_1, t_0)$ where u is given by (54a). However r is a comoving coordinate, so this is the coordinate label of the furthermost fundamental galaxy G that can influence R by means of any physical process taking place at or after time t_1 on G's world-line. Suppose now we let $t_1 \to 0$, where the origin of the universe is at $t = 0$ (i.e. $S(t_1) \to 0$ as $t_1 \to 0$). The integral may diverge; in that case, an observer at R can in principle see all the matter in the universe. This will be true for example in the Milne universe (48). However the integral may converge, and indeed by (49b) will do so in a FLRW universe filled with ordinary matter. In this case we say there is a *particle horizon* (Rindler 1956), and can then determine the coordinate value u_{ph} characterising the furthest galaxies that can possibly influence the event R at any time in its history:

$$u_{ph}(t_0) = u(0, t_0) = \int dt/S(t) \qquad (69a)$$

where the integral is from 0 to t_0. The corresponding physical scale at the present time is

$$d_{ph}(t_0) = S(t_0)\, u_{ph}(t_0), \qquad (69b)$$

giving the present-day distance to the set of galaxies separating those that can possibly have had causal connection with O at time t_0 from those that cannot. This is the largest present-day physical scale on which structures occuring can be explained in terms of physical processes acting causally since the creation of the universe. Their nature is particularly clarified by looking at causal diagrams for the FLRW universe representng the light cones at $\pm$ 45° (Penrose *1963, 1968,* Hawking and *Ellis 1973, Tipler et al, 1980). One can* think of this horizon as being defined by the set of particles at which the observed redshift (which increases with increasing distance from O) becomes infinite, because it corresponds to the particles for which $S(t_1) \to 0$. Two things are noteworthy: (a) the particle horizon always expands in the sense that once a galaxy G has entered the particle horizon of the observer O, it can never leave it (by (69a) $u_{ph}(t_0)$ is an increasing function of t_0 because $S(t)$ is *positive). Thus at later and later times, more* and more matter lies within the particle horizon of any observer O. (b) As long as the universe expands, the present physical scale (69b) must also be an increasing function of time.

The definition *given so far does not take into* account *limitations* on observations due to the early universe being opaque when $T/T_0 > 1000$ K, i.e. for $S(t)/S(t_0) < 1/1000$ (section 4.4). Let events P, Q lie on the furthest world lines O can see to by light emitted at the time of decoupling t_d when the universe becomes opaque; then the corresponding r-coordinate value is

$$u_{vh}(t_0) \;=\; u(t_d, t_0) \;=\; u(y_1, 1) \qquad (70a)$$

with u given by (54a) in the first case and (54b) in the second, where $y_1 = S(t_d)/S_0 = 1/1000$. The present day physical scale of this "visual horizon" is

$$d_{vh}(t_0) \;=\; S(t_0)\, u_{vh}(t_0); \qquad (70b)$$

it is the present-day distance to the furthest objects from which we can receive electromagnetic radiation (light, radio waves, and X-rays from more distant objects are scattered by the optically thick intervening medium). Isotropy of the microwave background radiation is evidence of homogeneity on this scale, because we measure such radiation from the surface of decoupling to have about the same temperature in all directions. It is important to notice then that at the time of decoupling the particle horizon occurred at the value

$$u_{pph} \;=\; u(0, t_d) \qquad (71a)$$

corresponding to a present-day physical distance of

$$d_{pph}(t_0) \;=\; S(t_0)\, u_{pph}\,. \qquad (71b)$$

This is the largest scale (evaluated at the present time) whose homogeneity can be explained via physical processes taking place in the early universe up to the time of decoupling. It follows from these definitions that

$$d_{ph}(t_0) = d_{vh}(t_0) + d_{pph}(t_0). \tag{72}$$

One can also evaluate these distances at the time t_d if desired, replacing $S(t_0)$ in (69b), (70b) and (71b) by $S(t_d)$, and again finding a relation equivalent to (72). By (56c) the present apparent angular size α_{pph} of the largest region that is causally connected at t_d is determined from the equation

$$d_{pph}(t_d) = S(t_d)\, u_{pph} = r_0(z_d)\, \alpha_{pph} \tag{73}$$

where $z_d \simeq 1000$ and $r_0(z)$ is determined from (57) if the universe is matter dominated since decoupling.

The relations so far hold for all equations of state. Now suppose the universe is radiation dominated at early times and matter dominated at late times; for simplicity we assume the change-over takes place at the time t_d (this may actually occur earlier or later, depending on the matter density, cf. section 4.2.2). Then initially (for $0 \leqq t \leqq t_d$)

$$S(t) = S_d\, (2H_d)^{1/2}\, t^{1/2}$$

(S_d, H_d constants) and the contribution to u during this era is

$$u_{pph} = 1/(H_d S_d).$$

Finally (for $t_d \leqq t \leqq t_0$)

$$S(t) = S_d\, \{(3/2)H_d(t-t_d) + 1\}^{2/3}$$

(the choice of constants ensuring continuity of S and $S^{\bullet}$ at $t = t_d$) and the contribution to u during this era is

$$u_{vh} = (2/H_d S_d)\, \{(S_0/S_d)^{1/2} - 1\}.$$

Evaluating the constants at $t = t_0$ shows

$$S_0 = S_d \{(3/2)H_d(t_0-t_d) + 1\}^{2/3}, \quad H_0 = H_d (S_d/S_0)^{3/2}$$

Thus $S_dH_d = S_0H_0 (S_0/S_d)^{1/2}$ and so

$$u_{pph} = 1/(S_0H_0) (S_0/S_d)^{-1/2} \tag{74a}$$

$$u_{vh} = 2/(S_0H_0) (S_0/S_d)^{-1/2} [(S_0/S_d)^{1/2} - 1] \tag{74b}$$

showing that

$$d_{vh}(t_0) = 2 [(S_0/S_d)^{1/2} - 1] d_{pph}(t_0)$$

which implies

$$d_{pph}(t_0) = 0.016 d_{vh}(t_0).$$

This is the "horizon problem": the scale of causal connection at the time of decoupling is much less than the size of the presently visible region of the universe (both sizes being evaluated at the present time t_0). Thus no causal process occuring after the creation of the universe can explain homogeneity on the scale of the visual horizon; but homogeneity on this scale is implied by the observed isotropy of the background radiation.

8.1 Limits to verification and prediction

Major limits in our ability to observationally verify the nature of the universe arise because of the limits imposed by the existence of particle horizons. In the $k = 0$ and $k = -1$ FLRW models with the standard topology, the implication is that we have seen an infinitely small fraction of the matter in the universe. Thus we can only determine the nature of conditions far away on the basis of philosophical assumptions which cannot be observationally verified or disproved (Ellis 1975, 1980). The question of what approach to take to these unobservable regions of the universe is puzzling. The standard approach by default implicitly assumes that conditions there are known, but no observational evidence backs up this assumption. In a $k = +1$ universe with standard topology the situation is different: we have then seen a finite fraction of all the matter in the universe, but there is much we cannot have seen. The only universes where we can have seen all the matter there is, are "small universes" (section 2.2.3 above; see also Ellis 1984b, Ellis and Schreiber 1986). Present observations do not disprove this possibility.

As a corollorary to this uncertainty, unless the universe is a "small universe" we are strictly only able to determine the structure of the universe in the past of the point R, and not anywhere to the future of its past light cone. The issue here is that objects we have not yet been able to see (and with which we have not yet had any causal contact) could generate gravitational waves that will invalidate any prediction we care to make about future events, e.g. an eclipse of the Moon or the return of Halley's comet. This has not happened in the past few hundred years, but this does not mean it cannot occur in the future. Thus we are only able to predict future events (such as that the Sun will rise tomorrow) on the basis of a "no-interference" condition: distant matter in the universe will not interfere with predictions we make on the basis of local knowledge. In effect this is an assumption of uniformity in parts of the universe we have not yet seen or had any causal connection with. A "cosmological principle" guaranteeing spatial homogeneity (and so validating the assumption that the universe is well-described by a FLRW model in regions for which we have no observational data) is such an assumption of uniformity, which is unverifiable.

The issue that arises is, how to deal with uncertainty in cosmology both in the verification of distant conditions and in attempts to predict to the future of our past light cone. Philosophical considerations inevitably play a dominant role in deciding on what approach to take to this issue.

9: LOCAL PHYSICS AND COSMOLOGY

9.1 Boundary conditions and local physics

The last section has started to emphasize that local physical conditions depend on the boundary conditions on physical fields in the universe (Bondi 1960, Ellis and Sciama 1972). Specifically, we noted there that local physical effects can only be predicted if distant conditions in the universe ("boundary conditions at infinity") are such as to not interfere with local predictions. In fact boundary conditions at the beginning of the universe govern what is possible in local physics. A particular example is "Olber's paradox": why is the sky dark at night ? (Bondi 1960, Harrison 1981), which is just a limit on the integrated background radiation from all sources, which in turn depends on initial conditions for the universe. If the background radiation temperature were over 300 K, life as we know it on earth would be impossible.

A further notable example is the arrow of time question: it seems probable that the local time-asymmetry of time-symmetric equations such as Maxwell's equations is due to time-asymmetric conditions for electromagnetic fields at the boundaries of the universe. This is intimately related to the expansion of the universe (introducing a fundamental time asymmetry into the universe and hence to its contents) which in turn may be related to the smoothness of initial conditions (Penrose 1979, 1981).

9.2 The Effect of Expansion on Local Physics

One way in which the time-asymmetry of the universe effects local physics is directly through its expansion. For example, Maxwell's equations for the electric and magnetic fields measured by fundamental observers in an expanding universe differ from their form in flat space-time (see e.g. Ellis 1973).

Of particular importance is the way the expansion of the universe affects geodesics, for these represent "free fall" motion; spherically symmetric test particles only deviate from geodesic motion if some force (other than gravity and inertia) acts on them. To illustrate this we consider briefly a particular problem relating to galaxy formation.

It is possible that voids in the distribution of galaxies arise from an explosion that takes place at an initial event at a time $t = t_1$, the effects of this explosion spreading out radially until a "freezing out" of this motion takes place at a time $t = t_2$, probably because the outwards moving matter encounters matter moving in the opposite direction originating at similar explosions at other places, and interacts with it to reduce the speed of both relative to comoving matter to nearly zero. We are interested in the length scale d of the resulting bubble measured at the present time t_0. The time t_2 precedes the present time t_0, and t_1 is after the time of decoupling t_d.

A model that represents limits on possible spreading of effects in an expanding universe is obtained by considering the explosion fragments moving freely, that is, following geodesics in the curved space-time. The fragments from the explosion have to catch up with matter moving away (due to the Hubble expansion in the expanding universe), and this limits the region which can be influenced. Geodesic motion represents this effect accurately. Clearly this model does not represent local physical effects, e.g. ram pressure resisting the motion, that will further limit the domain that can be affected by the explosions. Thus such free fall estimates give upper bounds on the scales that can be affected; local effects that do not feed energy into the particle motion will reduce the possible size of bubbles even further.

From (53) radial timelike geodesics of the metric (18b) obey the equations

θ = constant, ϕ = constant, and

$$dr/dt = E \,/\, \{S(t)\,(S^2(t) + E^2)^{1/2}\} \tag{81}$$

where E is a constant. Evaluating the speed of motion v_1 of the particle at the time t_1 relative to comoving observers, E is given by $E = v_1 S_1/(1 - v_1^2)^{1/2}$. Therefore if particles are scattered by an explosion at time t_1 at speed v_1, and they move freely (i.e. on geodesics) until time t_2 (either the present time t_0, or a time when they are brought to rest by meeting particles coming in the opposite direction) the change in their r–coordinate value is

$$u_{12} \;\equiv\; r_2 - r_1 \;=\; \int dt/\{S(t)\,(\gamma^2 S^2(t) + 1)^{1/2}\} \tag{76a}$$

with the integral taken from t_1 to t_2, where

$$\gamma^2 \;=\; (1/S_1^2)\,\{(1 - v_1^2)/v_1^2\}. \tag{76b}$$

The corresponding distance d evaluated at the time t_0 is given by

$$d = S(t_0)\, u_{12}\,. \tag{76c}$$

To evaluate these expressions we need an explicit form for S(t). For convenience we take it in the form

$$S(t) = \alpha(t - t_*)^n \tag{77}$$

where (a) n = 2/3 for k=0 dust, (b) n = 1/2 for k=0 radiation, or (c) n = 1 for the Milne universe. Then the Hubble constant H_0 is given by

$$H_0 \;\equiv\; (S^{\bullet}/S)|_0 \;=\; n/(t_0 - t_*) \tag{77a}$$

and so

$$S_0 \;=\; \alpha\,(n/H_0)^n. \tag{77b}$$

(we write $S(t_0) = S_0$, $S(t_1) = S_1$, $S(t_2) = S_2$) while the deceleration parameter q_0 is given by

$$q_0 \;\equiv -\,(S^{\bullet\bullet}/S)|_0\,(H_0)^{-2} \;=\; (1 - n)/n. \tag{77c}$$

It is not possible to obtain a general form for the integral when $S(t)$ is given by (77). However for large t_2, $u_{12} \sim \int dt/t^{2n}$, which converges for $n > 1/2$ and diverges if $n \leqq 1/2$. Thus in the low–density ($n = 1$) and high density ($n = 2/3$) (matter dominated) cases, the maximum comoving sphere that can be affected by such an explosion is strictly finite, no matter how long the particles continue to expand (cf Fairall 1987). In the radiation dominated high density case, on the other hand, arbitrarily large r–values can in principle be attained eventually. Furthermore, the local speed v of motion of the explosive effect relative to comoving matter is given by

$$v = S(t)\, dr/dt = E/(S^2(t)+E^2)^{1/2} \tag{78}$$

which always stays positive, but goes to zero as t goes to infinity and $S(t)$ increases indefinitely.

For the specific cases (77a–c) we find,

$$n = 1/2: \quad d = (c/2H_0\delta_1)\,(S_1/S_0)\,\{\text{arc cosh}(2\delta_2^2+1) - \text{arc cosh}(2\delta_1^2+1)\} \tag{79a}$$

where

$$\delta_1 \equiv (1 - v_1^2)^{1/2}/v_1, \qquad \delta_2 = (S_2/S_1)\delta_1; \tag{79b}$$

$$n = 2/3: \qquad d = (2c/H_0\delta_1^{1/2})\,(S_1/S_0)^{1/2} \int dz/(1 + z^4)^{1/2} \tag{79c}$$

where the integral is taken from $z_1 = \delta_1^{1/2}$ to $z_2 = \delta_2^{1/2}$; and

$$n = 1: \qquad d = (c/H_0)\,\{\text{arc sinh}\,(1/\delta_1) - \text{arc sinh}(1/\delta_2)\} \tag{79d}$$

In a general universe filled with pressure–free matter, $S(t)$ is determined by the Friedmann equation (43b) which shows

$$d = (c/H_0) \int \{(1-2q_0)y^2 + 2q_0 y\}^{-1/2}\,\{\delta_1^2(y/y_1)^2 + 1\}^{-1/2}\, dy \tag{80}$$

where the integral is taken from $y_1 = S_1/S_0 = (1+z_1)^{-1}$ to $y_2 = S_2/S_0 = (1+z_2)^{-1}$. Equations (79a) and (79c) above are respectively the special cases resulting when $q_0 = 1/2$ and $q_0 = 0$.

In *high density* cases (79a) and (79c), as z_1 is taken to higher and higher values (for fixed z_2 and v_1), d attains a maximum value and then declines: although the freely moving products of the explosion have longer and longer times to expand, in fact they are pulled back by the gravitational effect of the intervening matter, and for very large values of z_1 only very small distances are attained at z_2. For given z_2 and v_1, there is a value z_* of z_1 giving a maximum value of d, which is the largest distance (evaluated at t_0) that can be affected by an explosion ejecting products at speed v_1 at <u>any</u> value of z_1. The same effect occurs in the *low density cases* of pressure-free universes with $q_0 = 0.02$ and 0.05. Another way of thinking of the effect is that at earlier times, v_1 can be expected to be the same (because its local physical causes are unchanged) but the Hubble constant at that time is higher, so the speed has less effect; the explosion fragments are chasing matter which (at the same comoving distance) is moving away faster, and so is less successful in catching up with it. Smaller values of d will again be obtained for earlier galaxy formation, i.e. larger values of z_2, for each value of $z_1 > z_2$.

It is particularly interesting that the bubble size is in many cases insensitive to z_1 for quite a large range of values of z_1. The point is that there will in general be varied explosion times and strengths, so one might expect a large variation in resulting bubble sizes. The insensitivity to z_1 provides a possible explanation for relatively uniform sizes of observed bubbles; for example if $n = 2/3$, $v_1 = 10000$ km/sec and $z_2 = 4$, then roughly the same bubble size will result for z_1 anywhere between 10 and 45. Indeed these particular parameter values correspond reasonably well to the observed sizes of the foam-like structure.

Obviously if geodesic motion is determined by the expansion of the universe, geodesic deviation will also demonstrate such effects. A direct analysis (Ellis and Stoeger 1987a) shows this explicitly. For example, in an Einstein-de Sitter universe the motion of a freely falling particle relative to a (geodesically moving) fundamental observer is characterised by the deviation vector

$$\eta^a = C^a t^{2/3} + K^a t \qquad (75)$$

where C^a and K^a are covariantly constant along the world-lines. The implication is that a freely falling cloud of particles in an Einstein-de Sitter universe will feel the expansion of the universe, no matter how small the scale of the cloud. Consequently one might expect the expansion of the universe to imply a spiralling of planetary orbits when a mass such as the sun is imbedded in an

Einstein–de Sitter universe (cf Gautreau 1984). To the extent this does not happen, it is because the exactly smooth FLRW model is misleading about the nature of local physics in a realistic (lumpy) universe model. Thus one issue is, in applying the FLRW model to the real universe, what is the scale on which we expect Hubble's law to affect local physics ?

Closely related to the geodesic deviation equation is the use of Liouville's theorem to estimate the evolution of random velocities of galaxies.

9.3 Homogeneity and causality

An interesting question in relation to local physics in the expanding universe is, how do physical processes occur homogeneously ? As a specific example, consider nucleosynthesis. Because of the causal limitations implied by the past light cone, two distinct events P, Q on a surface of simultaneity $t = t_1$ do not share the same causal past, and indeed if far enough apart their causal pasts will be completely disjoint. Thus no common physical events can explain why nucleosynthesis takes place exactly simultaneously at these separate events.

The resolution is that physical processes along separate world lines are initially synchronized (in a FLRW universe) by the spatially homogeneous big bang; and thereafter, events on different world–lines remain synchronized, although they are causally disconnected, because in effect local physical processes act as perfect clocks along each separate world line. Thus spatial homogeneity, once initiated, is maintained by perfect time–keeping along the different world lines. This result will not hold if random processes (e.g. due to quantum fluctuations) occur; they will break this mechanism for maintaining homogeneity.

9.4 Deeper issues

There are other specific issues one can examine in relation to local physics in an expanding universe. However there is a deep underlying issue, namely the Machian problem: what is the relation of local physical laws to the nature of the universe ? If the universe were totally different, it is possible that local physical laws would be different; equivalently, in the cosmological context the division between local physical laws and boundary conditions may become meaningless. Such speculations have led to proposals such as Dirac's

Large Number's hypothesis, but the issue is still basically unresolved.

A related question is the relation between the nature of the universe and the existence of life. In essence, it is easy to construct universe models in which life is impossible (e.g. a recollapsing $k = +1$ universe where either the microwave background temperature never drops below 1000 K, or the total life-time of the universe is less than 10^7 years); so why is the real universe such as to admit the existence of intelligent life ? This takes one onto the hotly debated area of the *Anthropic Principle* (Barrow and Tipler 1986), suggesting that the universe must be such as to allow the existence of intelligent life. The proposal raises many intriguing questions but does not fully resolve them.

Overall, the point is that local physics is affected by the expansion and evolution of the universe, and the nature of this relation is not trivial and is not fully resolved. It is worth further exploration.

10: PROBLEMS AND POSSIBLE SOLUTIONS

10.1 Issues

10.1.1 "Best-buy" FLRW universe

The simplest approach is simply to look for a "best buy" FLRW universe. Despite the possible existence of dark matter of numerous kinds, the observational evidence is still as it was in 1974 and indicates, taken on its face value, a low-density universe. This "standard model" provides a good overall picture of the evidence presently available, and a satisfying overall picture of the evolution of matter in the expanding universe (see e.g. Longair 1985).

The problems arise when one starts to relate this to more realistic, "lumpy universe" models. On the one hand formation of inhomogeneous structures in this uniform background, and in particular galaxy formation, is still an unsolved issue. On the other hand the relation of these idealised, exactly homogeneous and isotropic universe models to reality is not clear. It is not obvious how particular galaxies or clusters relate to the comoving coordinates of the idealised model, nor how particular light rays correspond to the idealised geodesics of those models; indeed it is known that light rays in lumpy universes behave in a rather different way than in the smooth models. The standard approach to this issue is via examining perturbed FLRW models, which have not been discussed here (see particularly Bardeen 1980), and how they relate to possible observations of galaxies (Sasaki 1987) and background radiation (Sachs and Wolfe 1967). This has not been developed to the point of relating real observations to detailed perturbed universe models. An alternative is to consider the "fitting problem": can we prescribe a procedure determining the best-fit of a FLRW model to a lumpy universe by means of possible cosmological observations (Ellis and Stoeger 1987a) ? The approach is promising but needs development.

10.1.2 Relation to verification

Observational problems arise due to observational selection and detection effects on the one hand and source evolution on the other. More fundamentally, problems arise because we can only observe on the past light cone and (unless we live in a "small universe") the fraction of matter we can possibly see is strictly limited by particle horizons. Thus there are regions of the universe for which data is limited and others for which we cannot obtain any observational data whatever (Ellis 1975). An examination of the null data for cosmology, using observational coordinates, shows that the observationally determinable region is a FLRW universe if and only if the observational relations on the past null cone are precisely those predicted by FLRW models (Ellis et al 1985), but direct proof of homogeneity in this way is not practicable (Ellis 1980). Thus our belief in the applicability of the FLRW models results more from their overall coherence and general compatibility with the data than detailed observational confirmation of their nature.

10.1.3 Creation and the laws of physics

The big-bang origin of these models inevitably raises the issue of creation, and the special nature of creation leading to the observed universe. In particular, why are the laws of physics and the initial conditions in the universe such as to allow intelligent life ? This implies a very delicate balance of laws and environments.

The problem of "explaining" why the universe is the way it is, is compounded by the uniqueness of the universe: by definition, there is only one universe. This raises the complex issue of the relation of local physical laws to the universe: just as local physical laws affect the universe it is possible that the universe (which is the totality of all there is) affects local physical laws; indeed it is plausible that if the universe were very different these laws would also be different. Viewed differently, it may be that in the cosmological context, where initial conditions are given once and for all and cannot be altered, we cannot really make the usual distinction between physical laws and initial (or boundary) conditions.

10.1.4 Improbability

Until a decade ago, cosmologists were happy to accept that a Cosmological Principle (postulating either uniformity or simplicity, cf Ellis and Matravers 1985) "explained" why the real universe was very like a FLRW model. Recently there has been a move away from this view with various attempts to explain this uniformity in other ways. The basic problem here is that the FLRW models are very special, and so *a priori* are extremely unlikely within the general family of universe models. Why should the real universe have turned out to have such an extremely improbable geometry ?

10.2 Proposed solutions:

10.2.1 Chaotic cosmology

One approach is the "chaotic cosmology" program initiated by Misner (1969) trying to show that a wide variety of initial conditions would lead to *the smooth universe we see today. The initial version was only partially* satisfactory (MacCallum 1983). A recent version of this program is the inflationary universe model (Guth 1981, Gibbons et al 1983), where an early exponential expansion through many orders of magnitude smooths out the universe and also *solves the horizon problem: due to this expansion*, $d_{pph} >> d_{vh}$ (cf (74)). There are now many versions of this theory, but none are completely satisfactory.

10.2.2 Entropy considerations

Penrose (1979, 1981) argues that entropy considerations require of necessity that the universe be very smooth initially (but not finally), relating this issue to the arrow of time question. In effect this is a new version of the Cosmological Principle, tied in to the necessity that local physics proceed as we know it. However it is not clear that thermodynamic arguments apply in this context, and if so, there are questions as to the appropriate measure of gravitational entropy.

10.2.3 Small universes

This proposal (Ellis 1984b, Ellis and Schreiber 1986) suggests that very general inhomogeneous small universe models can provide the appearance of a FLRW universe. However their spatial topology cannot be the usual topology, and no proposal has yet been made as to how this topology, or the closure length scale, is determined.

10.2.4 Anthropic Principle

The strong version of this proposal (Barrow and Tipler 1986) is that the universe must admit intelligent life, perhaps for example because the nature of quantum mechanics requires observers to exist; it is then suggested that this must make the universe look the way it does. This kind of claim is intriguing but controversial. Its foundations (in terms of an ensemble of universes, the many worlds interpretation of quantum mechanics, or a much larger inhomogeneous universe) are open to debate. The weaker version (Carr and Rees 1979) simply enquires as to what conditions are necessary for intelligent life to exist. The hoped-for implication that the observed region of the universe must be like a FLRW model has not been fully substantiated.

10.2.5 Creation theory

The most ambitious project is a theory of creation of the universe, of necessity through some kind of quantum process (e.g. Hawking 1984). This involves specific approaches to quantum gravity and quantum cosmology, the foundations of both subjects being somewhat obscure. Interesting results obtained so far suggest this may go towards explaining present observations, but in terms of pre-existent structures whose origin is itself then open to question.

10.2.6 Fundamentals

In essence, it is probable there are four fundamental viewpoints on the origin of the present structures we see. These are, (1) Accident: conditions just happened initially, and so led to things being the way they are now, by chance. This appears to be very improbable in view of the high symmetry observed and the delicate balance required to allow intelligent life, i.e. the observed universe seems very unlikely within the set of all conceivable universes. However the application of probability arguments is dubious: the concept of probability cannot properly be applied to the universe itself, because it is unique (McCrea 1953). (2) Probability: although the structure of the universe appears very improbable, for various physical or other reasons it is in fact actually highly probable. (3) Necessity: coherence and consistency, as in the "bootstrap" approach to physics, require that things HAVE to be the way they are. The apparent alternatives are illusory. This is really just a strong version of the previous approach. (4) Design: the symmetries and delicate balances observed require an extraordinarily careful coherence of conditions and cooperation of effects, suggesting that in some sense they have been purposefully designed. Clearly this has theological implications.

It is possible elements of these different approaches could be combined in some form. A survey of the previous proposals will suggest that each of discussion within the domain of physics, rather than (1) or (4), which do not.

10.3 Conclusion

The standard models have a very clear geometrical and physical structure with great explanatory power, and represent satisfactorily much of what we see. However they need to be tested by comparison with a wider class of models with more general geometrical and physical properties, in order to see how good their explanatory power is relative to such alternatives. In looking at these alternatives it is important to distinguish what physical problems from metaphysical ones. Both kinds of problems are important in cosmology, but one must be clear which is which.

I thank D. R. Matravers for useful comments on this manuscript.

Appendix

The area distance relations (56–57) are of fundamental importance in relating the FLRW universe to observations. Three points are of interest. Firstly, there is a problem with the ($p = 0$, $\Lambda = 0$) relation (57a), namely it is unstable as $(q_0 z) \to 0$. Terrell (1977) derived a form of this relation, algebraically equivalent to (57a), that does not suffer from this problem. It is

$$r_0(z) = (z/H_0)\,(1+z)^{-2}\left\{1 + z(1-q_0)[1 + q_0 z + (1+2q_0 z+1)^{1/2}]^{-1}\right\}. \qquad (57e)$$

Secondly, the general formula for the area distance with p and Λ possibly non–zero is given by Dabrowski and Stelmach (1986). Thirdly, if either the source or the observer does not move with the fundamental 4–velocity then redshift changes according to (55b); the corresponding change in area distance is implied by results given by McKinley (1979).

REFERENCES

Audouze, J., 1980. In *Physical Cosmology*, ed. R. Balian et al. (Les Houches 1979). North Holland, Amsterdam.
Bahcall, J., and May, R.M., 1968. *Astrophys. Journ.* **152**, 37.
Bahcall, J., and Salpeter, E., 1965. *Astrophys. Journ.* **142**, 1677.
Balian, R., Audouze, J., and Schramm, D. N. (Eds), 1980. *Physical Cosmology* (Les Houches 1979). North Holland.
Bardeen, J.M., 1980. *Phys Rev* **D22**, 1882.
Barrow, J., Galloway, G. J., and Tipler, F. J., 1986. *Mon. Not. Roy. Ast. Soc.* **223**, 835.
Barrow, J., and Tipler, F., 1986. *The Anthropic Cosmological Principle.* Oxford University Press.
Bondi, H., 1960. *Cosmology*, Cambridge University Press.
Bonnor, W. B. and Ellis, G. F. R., 1986. *Mon. Not. Roy. Ast. Soc.* **218**, 605.
Bonometto, S. A., and Lucchin, F., 1978. *Astron. Astrophys.* **67**, 153.
Carr, B., and Rees, M. J., 1979. *Nature* **278**, 605.
Collins, C. B. and Ellis, G. F. R., 1979. *Phys. Reports* **56**, 65.
Dabrowski, M., and Stelmach, J., 1986. *Astron. Journ.* **92**, 1272.
Dicke, R. H., and Peebles, P. J. E., 1979. In *General Relativity*, Ed. S. W. Hawking and W. Israel (Cambridge University Press), 504.
Ehlers, J., Geren, P., and Sachs, R.K., 1968. *J Math Phys* , 1344.
Einstein, A., 1956. *The Meaning of Relativity.* Methuen.
Ellis, G. F. R., 1971a. In *General Relativity and Cosmology*, ed. R. K. Sachs, Proc. Int. School. Physics Enrico Fermi XLVII. (Academic Press, New York), 104.
Ellis, G. F. R., 1971b. *Gen. Rel. Grav.* **2**, 7.
Ellis, G. F. R., 1973. In *Cargese Lectures in Physics: Vol. 6*, Ed. E. Schatzmann. Gordon and Breach.
Ellis, G. F. R., 1975. *Qu. Jour. Roy. Ast. Soc.* **16**, 245.
Ellis, G. F. R., 1979. *Gen. Rel. Grav.* 11, 281.
Ellis, G. F. R., 1980. *Ann. New York Acad. Sci.* **336**, 130.
Ellis, G. F. R., 1984a. *Ann. Rev. Astron. Astrophys.* **22**, 157.
Ellis, G. F. R., 1984b. In *General relativity and Gravitation*, ed. B Bertotti et al, (Reidel), 215.
Ellis, G. F. R. and King, A. R., 1974. *Commun. Math. Phys.* **38**, 119.
Ellis, G. F. R., Maartens, R, and Nel, S D., 1978. *Mon. Not. Roy. Ast. Soc.* **184**, 439.
Ellis, G. F. R., and Matravers, D. R., 1985. In *A Random Walk in Relativity and Cosmology*, ed N. Dadhich, et al. Wiley Eastern, New Delhi.

Ellis, G. F. R., Matravers, D. R, and Treciokas, R., 1983. *Gen. Rel. Grav.* **15**, 931.
Ellis, G. F. R., Nel, S. D., Maartens, R., Stoeger, W., and Whitman, A. P., 1985. *Physics Reports* **124**, 315.
Ellis, G. F. R., Perry, J. J., and Sievers, A., 1984. *Astron. Journ.* **89**, 1124.
Ellis, G. F. R. and Sciama, D. W., 1972. In *General Relativity*, Ed. L. O'Raifeartaigh. Oxford University Press.
Ellis, G. F. R. and Schreiber, G., 1986. *Phys. Lett.* **115A**, 97.
Ellis, G. F. R. and Stoeger, W., 1987a. "The Fitting problem in cosmology", Preprint.
Ellis, G. F. R. and Stoeger, W., 1987b. "Horizons in Inflationary Universes". Preprint.
Ellis, G. F. R. and Tivon, G., 1985. *Observatory* **105**, 171.
Fairall, A. P., 1987. To appear, *Mon. Not. Ast. Soc. S. A.*
Field, G. B., Arp, H., and Bahcall, J., 1973. *The Redshift Controversy*. Benjamin.
Gautreau, R., 1984. *Phys. Rev.* **D29**, 198.
Gibbons, G. W., Hawking, S. W., and Siklos, S. T. C., 1983. *The Very Early Universe* (Cambridge University Press).
Goenner, H., and Muller-Hoissen, F., 1984. *Class. Qu. Grav.* **1**, 651.
Gott, J. R., Gunn, J. E., Schramm, D. N., and Tinsley, B. M., 1974. *Astrophys. Journ.* **194**, 543.
Gott, J. R., Gunn, J. E., Schramm, D. N., and Tinsley, B. M., 1976. *Scientific American*, March, 62.
Groth, E. J., and Peebles, P. J. E., 1977. *Astrophys. Journ.* **217**, 385.
Guth, A. H., 1981. *Phys. Rev.* **D23**, 347.
Harrison, E. R., 1967. *Mon. Not. Roy. Ast. Soc.* **137**, 69.
Harrison, E. R., 1977. *Vistas in Astronomy* **20**, 341.
Harrison, E. R., 1981. *Cosmology*. Cambridge University Press.
Harrison, E. R., and Noonan, T. W., 1979. *Astrophys. Journ.* **232**, 1979.
Hawking, S. W., 1984. In *General Relativity and Gravitation*, ed. B Bertotti et al. (Reidel), 491.
Hawking, S. W., and Ellis, G. F. R., 1968. *Astrophys. Journ.* **152**, 25.
Hawking, S. W., and Ellis, G. F. R., 1973. *The Large Scale Structure of Space-time*. Cambridge University Press.
Hiscox, W A, and Lindblom, L., 1987. "The problem of dissipation in relativistic fluids". Proceedings of 13th Texas Symposium on Relativistic Astrophysics.
Israel, W., and Stewart, J., 1979. *Ann. Phys. (NY)* **118**, 341.
Jones, J. E., 1974. *Proc. Roy. Soc. (Lond.)* **A340**, 263.
Kolb, E. W., Turner, M. S., Lindley, D., Olive, K., and Seckel, D., (Eds.)

1986. *Inner Space/Outer Space*. University of Chicago Press.
L'Apparent, V. de, Geller, M. J., and Huchra, J. P., 1986. *Astrophys. Journ.* **302**, L1.
Lightman, A. P., Press, W.H., Price, R.H., and Teukolsky, S.A, 1975. *Problem Book in Relativity and Gravitation*. Princeton University Press.
Loh, E. D., 1986. *Phys. Rev. Lett.* **57**, 2865.
Longair, M., 1985. *Observatory* **105**, 171.
MacCallum, M. A. H., 1983. In *The origin and evolution of galaxies*, Ed. B. J. T. Jones and J. E. Jones. (Reidel), 9.
Marx, G., and Sato, H., 1986. "Irreversibility of the cosmological expansion". Preprint, Kyoto University.
Matzner, R. A., 1986. *Publ. Ast. Soc. Pac.* **98**, 1049.
McCrea, W. H., 1953. *Rep. Prog. Phys.* **16**, 32.
McKinley, J.M., 1979. *Am. Journ. Phys.* **47**, 602.
Misner, C. W., 1969. *Phys. Rev. Lett.* **22**, 1071.
Narlikar, J., and Davidson, W., 1966. *Rep. Prog. Phys.,* XXIX, 539.
Peebles, P. J. E., 1971. *Physical Cosmology*. Princeton University Press.
Peebles, P. J. E., 1986. *Nature* **321**, 27.
Penny, A. J., and Dickens, R. J., 1984. Annual Report of South African Astronomical Observatory.
Penrose, R., 1963. In *Relativity Groups and Topology*, ed. C. M. de Witt and B. de Witt (Les Houches 1963). Gordon and Breach, New York.
Penrose, R., 1968. In *Battelle Rencontres*, ed C. M. de Witt and J. A. Wheeler. Benjamin, New York.
Penrose, R., 1979. In *General Relativity*, Ed. S. W. Hawking and W. Israel (Cambridge University Press), 581.
Penrose, R., 1981. In *Quantum Gravity 2*, ed. C. J. Isham, R. Penrose, and D. W. Sciama (Oxford University Press), 244.
Rees, M.J., 1980. In *Physical Cosmology*, Ed. Balian et al. North Holland.
Rindler, W., 1956. *Mon. Not. Roy. Ast. Soc.* **116**, 662.
Rindler, W., 1977. *Essential Relativity*. Springer Verlag.
Robertson, H. P., 1933. *Rev. Mod. Phys.* **5**, 62.
Rowan–Robinson, M., 1985. *The Cosmological Distance Scale*. Freeman, New York.
Sachs, R.K., and Ehlers, J., 1971. In *Astrophysics and General Relativity Vol 2*. Ed. M. Chretien. S. Deser and J. Goldstein. Gordon and Breach.
Sachs, R. K., and Wolfe, A. M., 1967. *Astrophys. Journ.* 147, 73.
Sandage, A. J., 1961. *Astrophys. Journ.* **133**, 355.
Sandage, A., 1982. *Astrophys. Journ.* **252**, 569.

Sandage, A., and Tamman, G., 1986. In *Iner Space/Outer Space*, ed. E. W. Kolb et al. University of Chicago Press.
Sasaki, M., 1987. "The magnitude--Redshift relation in a Perturbed Friedmann Universe". Preprint, Hiroshima University.
Schramm, D. N., 1983. *Physics Today*, April, 27.
Schroedinger, E., 1955. *Expanding Universes*, Cambridge University Press.
Schutz, B., 1985. *A First course in General Relativity*. Cambridge University Press.
Segal, I. E., 1976. *Mathematical Cosmology and Extragalctic Astronomy*. Academic Press, New York.
Setti, G., and van Hove, L., (Eds.) 1984. *Large Scale Structure of the Universe, Cosmology, and Fundamental Physics*. CERN, Geneva.
Stabell, R., and Refsdal, S. 1966. *Mon. Not. Roy. Ast. Soc.* **132**, 379.
Stoeger, W., Ellis, G. F. R., and Hellaby, C., 1987. To appear, *Mon. Not. Roy. Ast. Soc.*
Subramanian, K., and Chitre, S.M., 1987. *Astrophys. Journ.* **313**, 13.
Synge, J.L., and Schild, A., 1961. *Tensor Calculus*. Dover, New York.
Tipler, F. J., Clarke, C. J. S., and Ellis, G. F. R., 1980. In *General Relativity* ed A Held. Plenum Press.
Terrell, J., 1977. *Am. Journ. Phys.* **45**, 869.
Treciokas, R., and Ellis, G. F. R., 1971. *Commun. Math. Phys.* **23**, 1.
Turner, M. S., and Schramm, D. N., 1979. *Physics Today*, September, 42.
Weinberg, S., 1972. *Gravitation and Cosmology*. Wiley, New York.
Wilkinson, D.T., 1984. In *Large Scale Structure of the Universe, Cosmology and Fundamental Physics*, ed. G. Setti and L. Van Hove. CERN, Geneva.