10
The initial singularity in the universe

The expansion of the universe is in many ways similar to the collapse of a star, except that the sense of time is reversed. We shall show in this chapter that the conditions of theorems 2 and 3 seem to be satisfied, indicating that there was a singularity at the beginning of the present expansion phase of the universe, and we discuss the implications of space–time singularities.

In §10.1 we show that past-directed closed trapped surfaces exist if the microwave background radiation in the universe has been partially thermalized by scattering, or alternatively if the Copernican assumption holds, i.e. we do not occupy a special position in the universe. In §10.2 we discuss the possible nature of the singularity and the breakdown of physical theory which occurs there.

10.1 The expansion of the universe

In §9.1 we showed that many stars would eventually collapse and produce closed trapped surfaces. If one goes to a larger scale, one can view the expansion of the universe as the time reverse of a collapse. Thus one might expect that the conditions of theorem 2 would be satisfied in the reverse direction of time on a cosmological scale, providing that the universe is in some sense sufficiently symmetrical, and contains a sufficient amount of matter to give rise to closed trapped surfaces. We shall give two arguments to show that this indeed seems to be the case. Both arguments are based on the observations of the microwave background, but the assumptions made are rather different.

Observations of radio frequencies between 20 cm and 1 mm indicate that there is a background whose spectrum (shown in figure 62 (i)) seems to be very close to that of a black body at 2.7 °K (see, for example, Field (1969)). This background appears to be isotropic to within 0.2 % (figure 62 (ii); see, for example, Sciama (1971) and references given there for further discussion). The high degree of isotropy indicates that it cannot come from within our own galaxy (we

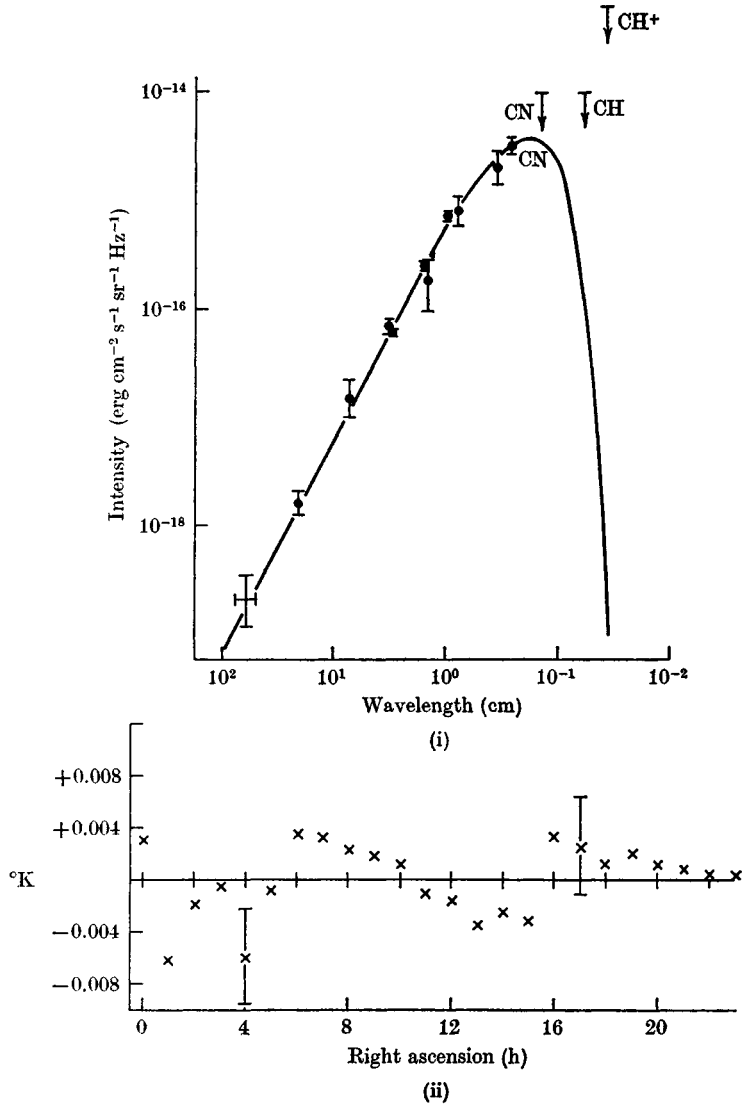

FIGURE 62

(i) The spectrum of the microwave background radiation. The plotted points show the observed values of the 'excess' background radiation. The solid line is a Planck spectrum corresponding to a temperature of 2.7 °K.

(ii) The isotropy of the microwave background radiation. The temperature distribution along the celestial equator is shown; more than two years of data have been averaged to obtain these points.

From D. W. Sciama, *Modern Cosmology*, Cambridge University Press, 1971.

are not symmetrically placed in the plane of the galaxy) but must be of extragalactic origin. At these frequencies we can see discrete sources some of whose distances are known from other evidence to be of the order of 10^{27} cm, so we know that the universe is transparent to this distance at these wavelengths. Thus radiation which is produced by sources at distances greater than 10^{27} cm must have propagated freely towards us for at least that distance.

Possible explanations of the origin of the radiation are:

(1) the radiation is black body radiation left over from a hot early stage of the universe;

(2) the radiation is the result of superposition of a very large number of very distant unresolved discrete sources;

(3) the radiation comes from intergalactic grains which thermalize other forms of radiation (perhaps infra-red).

Of these explanations, (1) seems the most plausible. (2) seems improbable, as there do not appear to be sufficient sources with the right sort of spectrum to produce an appreciable fraction of the observed radiation in this frequency range. Further, the small scale isotropy of the radiation implies that the number of discrete sources would have to be very large (of the order of the number of galaxies) and most galaxies do not seem to radiate appreciably in this region of the spectrum. (3) also seems unlikely, since the density of interstellar grains which would be needed is very large indeed. Although (1) seems the most probable, we will not base our arguments on it, since to do so would be to presuppose that the universe had a hot early stage.

The first argument involves the assumption of the Copernican principle, that we do not occupy a privileged position in space–time. We interpret this as implying that the microwave background radiation would appear equally isotropic to any observer whose velocity relative to nearby galaxies is small. In other words, we suppose there is an expanding timelike geodesic congruence (expanding because the galaxies are receding from each other, geodesic because they move under gravity alone with unit tangent vector V^a, say), representing the average motion of the galaxies, relative to which the microwave radiation appears almost isotropic. From the Copernican principle it also follows that most of the microwave background has propagated freely towards us from a very long distance ($\sim 3 \times 10^{27}$ cm). This is because the contribution to the background arising from a spherical shell of thickness $\mathrm{d}r$ and radius r about us will be approximately

independent of r, since the amount produced in the shell will be proportional to r^2 and the reduction of intensity due to distance will be inversely proportional to r^2. This will be the case until the redshift of the sources becomes appreciable, source evolution takes place, or curvature effects become significant. These effects will however only come in at a distance of the order of the Hubble radius, $\sim 10^{28}$ cm. Thus the bulk of the radiation will have travelled freely towards us from a distance $\gtrsim 10^{27}$ cm. From the fact that it remains isotropic travelling over such a long distance, we can conclude that on a large scale the metric of the universe is close to one of the Robertson–Walker metrics (§5.3). This follows from a result of Ehlers, Geren and Sachs (1968), which we will now describe.

The microwave radiation can be described by a distribution function $f(u, \mathbf{p})$ ($u \in \mathscr{M}$, $\mathbf{p} \in T_u$) defined on the null vectors in $T(\mathscr{M})$, which can be regarded as the phase space of the photons. If the distribution function $f(u, \mathbf{p})$ is exactly isotropic for an observer moving with four-velocity V^a, it will have the form $f(u, E)$ where $E \equiv -V^a p_a$. Since the radiation is freely propagating, f must obey the Liouville equation in $T(\mathscr{M})$. This states that f is constant along integral curves of the horizontal vector field $\mathbf{X}$, i.e. along any curve $(u(v), \mathbf{p}(v))$ where $u(v)$ is a null geodesic in $\mathscr{M}$ and $\mathbf{p} = \partial/\partial v$.

Because $f(u, E)$ is non-negative and must tend to zero as $E \to \infty$ (since otherwise the energy density of radiation would be infinite), there must be an open interval of E for which $\partial f/\partial E$ is non-zero. In this interval, one can express E as a function of f: $E = g(u, f)$. Then Liouville's equation implies that

$$\mathrm{d}E/\mathrm{d}v = g_{;a} p^a \tag{10.1}$$

on each null geodesic, where one regards g as a function on $\mathscr{M}$ with f fixed. Also,

$$\mathrm{d}E/\mathrm{d}v = -\mathrm{d}(V^a p_a)/\mathrm{d}v = -V_{a;b} p^a p^b. \tag{10.2}$$

One can decompose p^a into a part along V^a and a part orthogonal to V^a: $p^a = E(V^a + W^a)$, where $W^a W_a = 1$, $W^a V_a = 0$. Then from (10.1) and (10.2),

$$\mathrm{d}g/\mathrm{d}t + \tfrac{1}{3}\theta g + (g\dot{V}_a + g_{;a}) W^a + g\sigma_{ab} W^a W^b = 0$$

holds for all unit vectors W^a orthogonal to V^a, where $\mathrm{d}g/\mathrm{d}t$ is the rate of change of g along the integral curves of $\mathbf{V}$. Separating out spherical harmonics,

$$\sigma_{ab} = 0, \tag{10.3a}$$

$$\dot{V}_a + (\log g)_{;a} = \alpha V_a, \tag{10.3b}$$

$$\tfrac{1}{3}\theta = -\mathrm{d}(\log g)/\mathrm{d}t. \tag{10.3c}$$

Since we assumed that $\dot{V}_a$ was zero, (10.3*b*) shows that V_a is orthogonal to the surfaces $\{g = \text{constant}\}$, and this implies that the vorticity ω_{ab} is zero. As $\dot{V}^a = 0$, $V_{[a,b]} = 0$. Thus one can write V_a as the gradient of a function t: $V_a = -t_{,a}$.

The energy–momentum tensor of the radiation will have the form

$$T_{ab} = \tfrac{4}{3}\mu_r V_a V_b + \tfrac{1}{3}\mu_r g_{ab},$$

where $\mu_r = \int fE^3 \, dE$. Since the motion of the galaxies relative to the integral curves of V^a is small, their contribution to the energy–momentum tensor can be approximated by a smooth fluid with density μ_G, four-velocity V_a and negligible pressure. It now follows that the geometry of the space–time is the same as that of a Robertson–Walker model. To see this, note that

$$\begin{aligned}(V^a{}_{;b})_{;a} &= \tfrac{1}{3}(\theta(\delta^a{}_b + V^a V_b))_{;a} \\ &= (V^a{}_{;a})_{;b} + R^{ca}{}_{ba} V_c = \theta_{;b} + R_{ba} V^a.\end{aligned}$$

Multiplying this equation by $h^b{}_c = g^b{}_c + V^b V_c$, one finds

$$h^{bc} R_{ca} V^a = -\tfrac{2}{3} h^{bc} \theta_{;c}.$$

The left-hand side vanishes by the field equations. Thus θ is constant on the surfaces of constant t (which are also the surfaces of constant g). One can define a function $S(t)$ from θ by $S^{\cdot}/S = \frac{1}{3}\theta$; then the Raychaudhuri equation (4.26) takes the form

$$3S^{\cdot\cdot}/S + 4\pi\mu - \Lambda = 0,$$

which implies that $\mu = \mu_G + 2\mu_R$ is also constant on the surfaces $\{t = \text{constant}\}$. From the definition of μ_R we see that the terms μ_G and μ_R are separately constant on these surfaces.

The trace-free part of (4.27) shows that $C_{abcd} V^b V^d = 0$. The Gauss–Codacci equations (§2.7) now give for the Ricci tensor of the three-spaces $\{t = \text{constant}\}$ the formula

$$\begin{aligned}{}^3R_{ab} &= h_a{}^c h_b{}^d R_{cd} + R_{acbd} V^c V^d + \theta\theta_{ab} + \theta_{ac}\theta^c{}_b \\ &= 2h_{ab}(-\tfrac{1}{3}\theta^2 + 8\pi\mu + \Lambda).\end{aligned}$$

However for a three-dimensional manifold, the Riemann tensor is completely determined by the Ricci tensor, as

$${}^3R_{abcd} = \eta_{ab}{}^{ef}(-{}^3R_{ef} + \tfrac{1}{2}{}^3R h_{ef})\eta^f{}_{cd}.$$

This shows that each three-space $\{t = \text{constant}\}$ is a three-space of constant curvature $K(t) = \frac{1}{3}(8\pi\mu + \Lambda - \frac{1}{3}\theta^2)$. Integrating the Raychaudhuri equation shows that

$$K(t) = \tfrac{1}{3}(8\pi\mu + \Lambda - 3S^{\cdot 2}/S^2) = k/S^2, \tag{10.4}$$

where k is a constant. By normalizing S, one can set $k = +1, 0$ or -1. The four-dimensional space–time manifold is the orthogonal product of these three-spaces and the t-line. Thus the metric can be written in comoving coordinates as

$$ds^2 = -dt^2 + S^2(t)\, d\gamma^2,$$

where $d\gamma^2$ is the metric of a three-space of constant curvature k. But this is just the metric of a Robertson–Walker space (see §5.3).

We shall now show that in any Robertson–Walker space containing matter with positive energy density and $\Lambda = 0$ there is a closed trapped surface lying in any surface $\{t = \text{constant}\}$. To see this, we express $d\gamma^2$ in the form

$$d\gamma^2 = d\chi^2 + f^2(\chi)\,(d\theta^2 + \sin^2\theta\, d\phi^2)$$

where $f(\chi) = \sin\chi$, χ or $\sinh\chi$ if $k = +1$, 0 or -1 respectively. Consider a two-sphere $\mathscr{T}$ of radius χ_0 lying in the surface $t = t_0$. The two families of past-directed null geodesics orthogonal to $\mathscr{T}$ will intersect the surfaces $\{t = \text{constant}\}$ in two two-spheres of radius

$$\chi = \chi_0 \pm \int_{t_0}^{t} dt/S(t). \tag{10.5}$$

The surface area of a two-sphere of radius χ is $4\pi S^2(t) f^2(\chi)$. Thus both families of null geodesics will be converging into the past if, at $t = t_0$,

$$\frac{d}{dt}(S^2(t) f^2(\chi)) > 0$$

holds for both values of χ given by (10.5). This will be the case if

$$\frac{\dot{S}(t_0)}{S(t_0)} > \pm \frac{f'(\chi_0)}{S(t_0) f(\chi_0)}.$$

But by (10.4), this holds if

$$(\tfrac{8}{3}\pi\mu(t_0)\, S^2(t_0) - k)^{\frac{1}{2}} > \pm f'(\chi_0)/f(\chi_0).$$

This will be the case if $S(t_0)\,\chi_0$ is taken to be greater than $\sqrt{(3/8\pi\mu_0)}$ for $k = 0$ or -1, and to be greater than $\min(\sqrt{(3/8\pi\mu_0)}, \frac{1}{2}\pi)$ if $k = +1$.

An intuitive way of viewing this result is that at time t_0 a sphere of coordinate radius χ_0 will contain a mass of the order of $\frac{4}{3}\pi\mu_0 S^3(t_0)\,\chi_0{}^3$, and so will be within its Schwarzschild radius if $S(t_0)\,\chi_0$ is less than $\frac{8}{3}\pi\mu_0 S(t_0)^3\chi_0{}^3$, i.e. if $S(t_0)\,\chi_0$ is greater than the order of $\sqrt{(3/8\pi\mu_0)}$. We shall call $\sqrt{(3/8\pi\mu_0)}$ the *Schwarzschild length* of matter density μ_0.

So far, we have assumed the microwave radiation is exactly isotropic. This is of course not the case; and this corresponds to the fact

that the universe is not exactly a Robertson–Walker space. However, the large scale structure of the universe should be close to that of a Robertson–Walker model, at least back to the time when the radiation was emitted or last scattered. (One can in fact use the deviations of the microwave radiation from exact isotropy to estimate how large the departures from a Robertson–Walker universe are.) For a sufficiently large sphere, the existence of local irregularities should not significantly affect the amount of matter in the sphere, and hence should not affect the existence of a closed trapped surface round us at the present time.

The above argument did not depend on the spectrum of the microwave radiation, but it did involve the assumption of the Copernican principle. The argument we shall now give does not involve the Copernican principle, but does to a certain extent depend on the shape of the spectrum. We shall assume that the approximately black body nature of the spectrum and the high degree of small scale isotropy of the radiation indicate that it has been at least partially thermalized by repeated scattering. In other words, there must be enough matter on each past-directed null geodesic from us to cause the opacity to be high in that direction. We shall now show that this matter will be sufficient to make our past light cone reconverge.

Consider a point p representing us at the present time, and let W^a be a past-directed unit vector parallel to our four-velocity.

The affine parameter v on the past-directed null geodesics through p may be normalized by $K^a W_a = -1$, where $\mathbf{K} = \partial/\partial v$ is the tangent vector to the null geodesics. The expansion $\hat{\theta}$ of these null geodesics will obey (4.35) with $\hat{\omega} = 0$. Thus, providing $R_{ab} K^a K^b \geqslant 0$, $\hat{\theta}$ will be less than $2/v$. It follows that at $v = v_1 > v_0$,

$$\int_{v_0}^{v_1} R_{ab} K^a K^b \, \mathrm{d}v - 2/v_0 > \hat{\theta},$$

so $\hat{\theta}$ will become negative if there is some v_0 such that

$$\int_{v_0}^{v_1} R_{ab} K^a K^b \, \mathrm{d}v > 2/v_0.$$

Using the field equations with $\Lambda = 0$, this becomes

$$\tfrac{1}{2} v_0 \int_{v_0}^{v_1} 8\pi T_{ab} K^a K^b \, \mathrm{d}v > 1. \tag{10.6}$$

At centimetre wavelengths, the largest ratio of opacity to density for matter at reasonable densities is that given by Thomson scattering off

free electrons in ionized hydrogen. Thus the optical depth to a distance v will be less than

$$\int_0^v \kappa\rho(K^aV_a)\,\mathrm{d}v,$$

where κ is the Thomson scattering opacity per unit mass, ρ is the density of the matter, and V_a is the local velocity of the gas. The redshift z of the matter is given by $z = K^aV_a - 1$. Since no matter has been seen with significant blue-shifts, we shall assume K^aV_a is always greater than one on our past light cone, out to an optical depth unity. As galaxies are observed at these wavelengths with redshifts of 0.3, most of the scattering must occur at redshifts greater than this. (In fact if quasars really are cosmological, the scattering must occur at redshifts greater than two.) With a Hubble constant of 100 Km/sec/Mpc ($\sim 10^{10}$ years^{-1}), a redshift of 0.3 corresponds to a distance of about 3×10^{27} cm. Taking this value for v_0, the contribution to the integral (9.9) of the matter causing the scattering is

$$3.7\times10^{28}\int_{v_0}^{v_1}\rho(K_aV^a)^2\,\mathrm{d}v,$$

while the optical depth of the matter between v_0 and v_1 is less than

$$6.6\times10^{27}\int_{v_0}^{v_1}\rho(K^aV_a)\,\mathrm{d}v.$$

Since $K^aV_a \geqslant 1$, it can be seen that the inequality (10.6) will be satisfied at an optical depth of less than 0.2. If the optical depth of the universe was less than 1, one would not expect either an almost black body spectrum or such a high degree of small scale isotropy, unless there was a very large number of discrete sources which covered only a small fraction of the sky and each of which had a spectrum roughly the same as a 3 °K black body but with much higher intensity. This seems rather unlikely. Thus we believe that the condition (4) (iii) of theorem 2 is satisfied, and so there should be a singularity somewhere in the universe provided the other conditions hold.

Because of its generality, theorem 2 does not tell us whether the singularity will be in our past or in the future of our past. Although it might seem obvious that the singularity should be in our past, one can construct an example in which it is in the future: consider a Robertson–Walker universe with $k = +1$ which collapses to a singularity at some time $t = t_0$, and which asymptotically approaches an Einstein static universe for $t\to-\infty$. This satisfies the energy assumption, and contains points whose past light cones start reconverging (because they

meet up around the back). However the singularity is in the future. Of course this is a rather unreasonable example but it shows that one has to be careful. We shall therefore give an argument based on theorem 3 which indicates that the universe contains a singularity in our past, providing that the Copernican principle holds. Theorem 3 is similar to theorem 2, but requires that all the past-directed timelike geodesics from a point shall start to reconverge, instead of all the null geodesics. This condition is not satisfied in the example given above, though it is there satisfied by the future-directed geodesics from any point.

By an argument similar to that given above for the null geodesics, the convergence $\theta(s)$ of the past-directed timelike geodesics from a point p will be less than

$$\frac{3}{s_0} - \int_{s_0}^{s} R_{ab} V^a V^b \, \mathrm{d}s,$$

where s is proper distance along the geodesics, $\mathbf{V} = \partial/\partial s$ and $s > s_0$. Let $\mathbf{W}$ be a past-directed timelike unit vector at p, and let $c \equiv -V^a W_a|_p$ (so $c \geqslant 1$). Then θ will become less than $-c$ within a distance R_1/c along any geodesic if there is some R_0, $R_1 > R_0 > 0$, such that

$$\int_{R_0/c}^{R_1/c} R_{ab} V^a V^b \, \mathrm{d}s > c(3/R_0 + \epsilon) \tag{10.7}$$

along that geodesic. Condition (3) of theorem 3 will then be satisfied with $b = \max(R_1, (3\epsilon)^{-1})$.

To make (10.7) appear more similar to (10.6), we shall introduce an affine parameter $v = s/c$ along the timelike geodesics; then (10.7) becomes

$$\tfrac{1}{3}R_0 \int_{R_0}^{R_1} R_{ab} K^a K^b \, \mathrm{d}v > 1 + \tfrac{1}{3}R_0\epsilon, \tag{10.8}$$

where $\mathbf{K} = \partial/\partial v$ and $K^a W_a|_p = -1$. We cannot verify this condition directly by observation as in the case of (10.6) because it refers to timelike geodesics. We therefore have to appeal to the arguments given in the first part of this section to show that the universe is close to a Robertson–Walker universe model at least back to the time the microwave background radiation was last scattered.

In a Robertson–Walker model, let $\mathbf{W}$ be the vector $-\partial/\partial t$. Along a past-directed timelike geodesic through p,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}v}(W_a K^a) &= W_{a;b} K^a K^b \\ &= -\frac{1}{S}\frac{\mathrm{d}S}{\mathrm{d}t}\{(W^a K_a)^2 - 1/c^2\}. \end{aligned}$$

Therefore, providing that $\mathrm{d}S/\mathrm{d}t > 0$, $W_a K^a \leqslant -1$. However

$$W^a K_a = \mathrm{d}t/\mathrm{d}v;$$

thus for some $\epsilon > 0$, (10.8) will be satisfied for every geodesic provided that there are times t_2, t_3 with $t_2 < t_3 < t_p$ such that

$$\frac{t_p - t_3}{3} \int_{t_2}^{t_3} R_{ab} K^a K^b (-W_c K^c)^{-1} \, \mathrm{d}t > 1. \tag{10.9}$$

By the field equations with $\Lambda = 0$,

$$R_{ab} K^a K^b = 8\pi\{(\mu + p)(W_a K^a)^2 - \tfrac{1}{2}(\mu - p)c^{-2}\}.$$

Therefore, providing $p \geqslant 0$,

$$R_{ab} K^a K^b \geqslant 4\pi\mu (W_a K^a)^2.$$

Thus (10.9) will be satisfied if

$$\frac{t_p - t_3}{3} \int_{t_2}^{t_3} 4\pi\mu \, \mathrm{d}t > 1. \tag{10.10}$$

Assuming that the microwave radiation has a black body spectrum at 2.7 °K, its energy density is about 10^{-34} gm cm^{-3} at the present time. If this radiation is primaeval, its energy density will be proportional to S^{-4}. Since $S^{-1} = O(t^{\frac{1}{2}})$ as t tends to zero, one can see that (10.10) can be satisfied by taking t_3 to be $\frac{1}{2}t_p$ and t_2 to be sufficiently small. How small t_2 has to be depends on the detailed behaviour of S, which in turn depends on the density of matter in the universe. This is somewhat uncertain, but seems to lie between 10^{-31} gm cm^{-3} and 5×10^{-29} gm cm^{-3}. In the former case, t_2 will have to be such that $S(t_p)/S(t_2) \geqslant 30$, and in the latter case, $S(t_p)/S(t_2) \geqslant 300$. Since the microwave radiation seems to be all pervasive, any past-directed timelike geodesic must pass through it. Thus an estimate based on the Robertson–Walker models should be a good approximation for its contribution to (10.10), provided that the radiation was not emitted more recently than t_2, and provided that a Robertson–Walker model is a good approximation back that far. From the arguments at the beginning of this section, the latter should be the case provided that the radiation has propagated freely towards us since t_2. However there may be ionized intergalactic gas present with a density as high as 5×10^{-29} gm cm^{-3}, in which case the radiation could be last scattered at a time t such that $S(t_p)/S(t) \sim 5$. The optical depth back to a time t is

$$\int_t^{t_p} \kappa \mu_{\mathrm{gas}} \, \mathrm{d}t, \tag{10.11}$$

where κ is at most 0.5 if μ is measured in gm cm^{-3} and t in cm.

As before, there can be no significant opacity back to $t = t_p - 10^{17}$ sec, since we see objects at distances of at least 3×10^{27} cm. Taking t_3 to have this value, we see that the gas density will cause (10.11) to be satisfied for a value of t_2 corresponding to an optical depth of at most 0.5.

Thus the position is as follows. We assume the Copernican principle, and that the microwave radiation has been emitted either before a time t_2 such that $S(t_p)/S(t_2) \approx 300$, or before the time corresponding to the optical depth of the universe being unity, if this is less than t_2. In the former case, condition (2) of theorem 3 will be satisfied by the radiation density, and in the latter case by the gas density. Thus if the usual energy conditions and causality conditions hold, we can conclude that there should be a singularity in our past (i.e. there should be a past-directed non-spacelike geodesic from us which is incomplete).

Suppose one takes a spacelike surface which intersects our past light cone and takes a number of points on that surface; can one say that there is a singularity in each of their pasts? This will be the case if the universe is sufficiently homogeneous and isotropic in the past to converge all the past-directed timelike geodesics from these points. In view of the close connection between the convergence of timelike geodesics and closed trapped surfaces, we would expect this to be the case if the universe is homogeneous and isotropic at that time on the scale of the Schwarzschild length $(3/8\pi\mu)^{\frac{1}{2}}$.

We have direct evidence of the homogeneity of the universe in our past from the measurements of Penzias, Schraml and Wilson (1969), who found that the intensity of the microwave background is isotropic to within 4 % for a beam width of 1.4×10^{-3} square degrees. Assuming that the microwave radiation has not been emitted since a surface in our past corresponding to optical depth unity, the observed intensity will be proportional to $T^4/(1+z)^4$ where T is the effective temperature of the observed point on the surface and z is its redshift. Variations in the observed intensity can arise in four ways:

(1) by a Doppler shift caused by our own motion relative to the black body radiation (Sciama (1967), Stewart and Sciama (1967));

(2) by variations in the gravitational redshift caused by inhomogeneities in the distribution of matter between us and the surface (Sachs and Wolfe (1967), Rees and Sciama (1968));

(3) by Doppler shifts caused by local velocity disturbances of the matter at the surface; and

(4) by variations of the effective temperature of the surface.

(In fact the division between (1), (2) and (3) depends on the standard of reference and has heuristic value only.) Thus the observations indicate that irregularities in the temperature with an angular size of 3′ of arc have relative amplitudes of less than 1 %, and that there are no local fluctuations of the velocity of the matter, on the same scale, of greater than 1 % of the velocity of light. A region on the surface which had an angular diameter 3′ of arc would correspond to a region which had a diameter now of about 10^7 light years. If the surface of optical depth unity is at a redshift of about 1000 (this is the most it could be), the Schwarzschild length at that time would correspond to a region whose present diameter was about 3×10^8 light years. Thus it would seem that every point on the surface of optical depth unity should have a singularity in its past.

More indirect evidence on the degree of homogeneity of the universe in the early stages comes from the fact that observations of the helium content of a number of objects agree with calculations of helium production by Peebles (1966), and Wagoner, Fowler and Hoyle (1968), who assumed the universe was homogeneous and isotropic at least back to a temperature of about 10^9 °K. On the other hand calculations of anisotropic models have shown that in these models very different amounts of helium are produced. Thus if one accepts that there is a fairly uniform density of helium in the universe (there are some doubts about this), and that this helium was produced in the early stages of the universe, one can conclude that the universe was effectively isotropic and hence homogeneous when the temperature was 10^9 °K. One would therefore expect a singularity to occur in the past of each point at this time.

Misner (1968) has shown that if the temperature reaches 2×10^{10} °K a large viscosity arises from collisions between electrons and neutrinos. This viscosity would damp out inhomogeneities whose lengths correspond to present values of 100 light years, and reduce anisotropy to a comparatively small value. Thus if one accepts this as the explanation for the present isotropy of the universe (and it is a very attractive one), one would conclude that there should be a singularity in the past of every point when the temperature was about 10^{10} °K.

10.2 The nature and implications of singularities

One might hope to learn something about the nature of the singularities that are likely to occur by studying exact solutions with

singularities. However although we have shown that the *occurrence* of a singularity is not prevented by small perturbations of the initial conditions, it is not clear that the *nature* of the singularity which occurs will be similarly stable. Although we have shown in §7.5 that the Cauchy problem is stable under small perturbations of the initial conditions, this stability applies only to compact regions of the Cauchy development, and a region containing a singularity is non-compact unless the singularity corresponds to imprisoned incompleteness. In fact we can give an example where the nature of the singularity is not stable. Consider a uniform spherically symmetric cloud of dust collapsing to a singularity. The metric inside the dust will be similar to that of part of a Robertson–Walker universe, while that outside will be the Schwarzschild metric. Both inside and outside the dust, the singularity will be spacelike (figure 63 (i)). Suppose now one adds a small electric charge density to the dust. The metric outside the dust now becomes part of the Reissner–Nordström solution for $e^2 < m^2$ (figure 63 (ii)). There will be a singularity inside the dust, as a sufficiently small charge density will not prevent the occurrence of infinite density. The nature of the singularity inside the dust will presumably depend on the charge distribution. However the important point is that once the surface of the dust has passed a point p inside $r = r_+$, whatever happens inside the dust cannot affect the portion sq of the timelike singularity.

If one now increases the charge density so that it becomes greater than the matter density, it is possible for the cloud to pass through the two horizons at $r = r_+$ and $r = r_-$ and to re-expand into another universe without any singularity occurring inside the dust, although there is a timelike singularity outside the dust (J. M. Bardeen, unpublished), as indeed there ought to be by theorem 2 (see figure 63 (iii)).

This example is very important as it shows that there can be timelike singularities, that the matter can avoid hitting the singularities, and that it can pass through a 'wormhole' into another region of space–time or into another part of the same space–time region. Of course one would not expect to have such a charge density on a collapsing star, but since the Kerr solution is so similar to the Reissner–Nordström solution one might expect that angular momentum could produce a similar wormhole. One might speculate therefore that prior to the present expansion phase of the universe there was a contraction phase in which local inhomogeneities grew large and isolated singu-

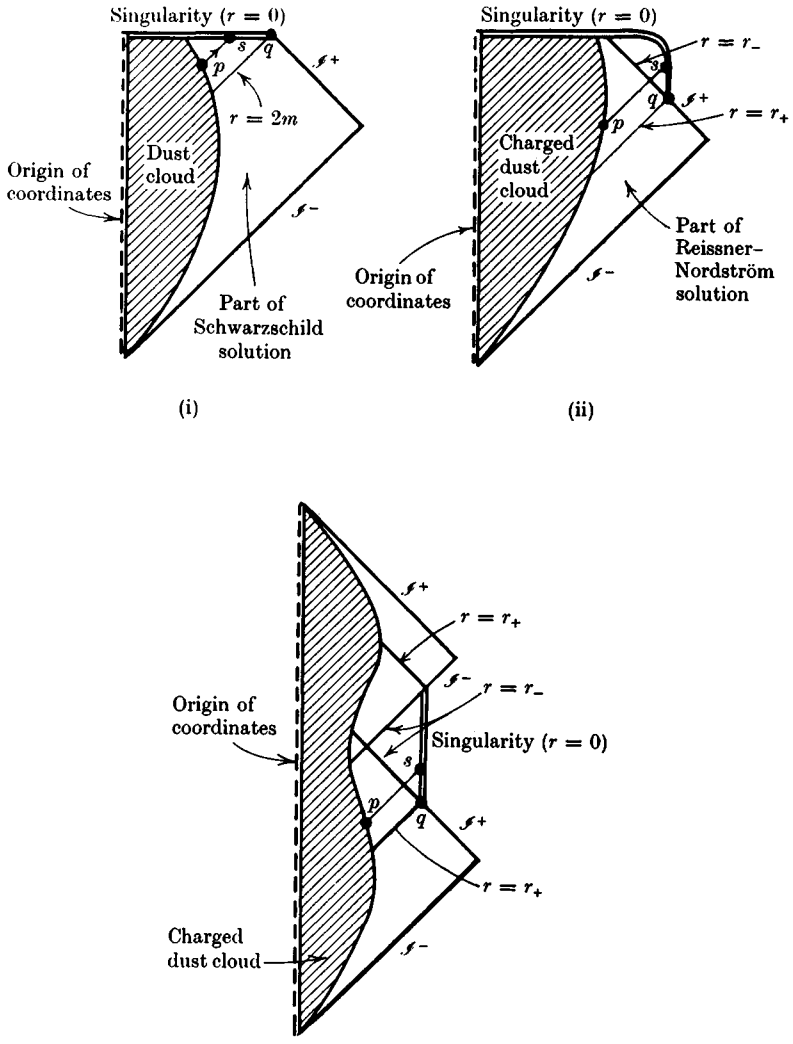

FIGURE 63

(i) Collapse of a spherical dust cloud.

(ii) Collapse of a charged dust cloud, where the charge is too small to prevent the occurrence of a singularity in the dust.

(iii) Collapse of a charged dust cloud, where the charge is large enough to prevent the occurrence of a singularity in the dust cloud; the singularity occurs outside the dust, which bounces and re-expands into a second asymptotically flat space.

larities occurred, most of the matter avoiding the singularities and re-expanding to give the present observed universe.

The fact that singularities must occur within the past of every point at an early time when the density was high, places limits on the separation of the singularities. It might be that the set of geodesics which hit these singularities (i.e. which are incomplete) was a set of measure zero. Then one might argue that the singularities would be physically insignificant. However this would not be the case because the existence of such singularities would produce a Cauchy horizon and hence a breakdown of one's ability to predict the future. In fact this could provide a way of overcoming the entropy problem in an oscillating world model since at each cycle the singularity could inject negative entropy.

So far, we have been exploring the mathematical consequences of taking a Lorentz manifold as the model for space–time, and requiring that the Einstein field equations (with $\Lambda = 0$) hold. We have shown that according to this theory, there should be singularities in our past associated with the collapse of the universe, and singularities in the future associated with the collapse of stars. If Λ is negative, the above conclusions would be unaffected. If Λ is positive, observations of the rate of change of expansion of the universe (Sandage, (1961, 1968)) indicate that Λ cannot be greater than $3 \times 10^{-55}\,\mathrm{cm}^{-2}$. This is equivalent to a negative energy density of $3 \times 10^{-27}\,\mathrm{gm\,cm}^{-3}$. Such a value of Λ could have an effect on the expansion of the whole universe, but it would be completely swamped by the positive matter density in a collapsing star. Thus it does not seem that a Λ term can enable us to avoid facing the problem of singularities.

It may be that General Relativity does not provide a correct description of the universe. So far it has only been tested in situations in which departures from flat space are very small (radii of curvature of the order of 10^{12} cm). Thus it is a tremendous extrapolation to apply it to situations like collapsing stars where the radius of curvature becomes less than 10^{6} cm. On the other hand the theorems on singularities did not depend on the full Einstein equations but only on the property that $R_{ab}K^aK^b$ was non-negative for any non-spacelike vector K^a; thus they would apply also to any modification of General Relativity (such as the Brans–Dicke theory) in which gravity is always attractive.

It seems to be a good principle that the prediction of a singularity by a physical theory indicates that the theory has broken down, i.e. it

no longer provides a correct description of observations. The question is: when does General Relativity break down? One would expect it to break down anyway when quantum gravitational effects become important; from dimensional arguments it seems that this should not happen until the radius of curvature becomes of the order of 10^{-33} cm. This would correspond to a density of 10^{94} gm cm^{-3}. However one might question whether a Lorentz manifold is an appropriate model for space–time on length scales of this order. So far experiments have shown that assuming a manifold structure for lengths greater than 10^{-15} cm gives predictions in agreement with observations (Foley *et al.* (1967)), but it may be that a breakdown occurs for lengths between 10^{-15} and 10^{-33} cm. A radius of 10^{-15} cm corresponds to a density of 10^{58} gm cm^{-3} which for all practical purposes could be regarded as a singularity. Thus maybe one should construct a surface by Schmidt's procedure (§ 8.3) around regions where the radius of curvature is less than, say, 10^{-15} cm. On our side of this surface a manifold picture of space–time would be appropriate, but on the other side an as yet unknown quantum description would be necessary. Matter crossing the surface could be thought of as entering or leaving the universe, and there would be no reason why that entering should balance that leaving.

In any case, the singularity theorems indicate that the General Theory of Relativity predicts that gravitational fields should become extremely large. That this happened in the past is supported by the existence and black body character of the microwave background radiation, since this suggests that the universe had a very hot dense early phase.

The theorems on the existence of singularities could possibly be refined somewhat, but on our view they are already adequate. However they tell us very little about the nature of the singularities. One would like to know what kind of singularities could occur in generic situations in General Relativity. A possible way of approaching this would be to refine the power series expansion technique of Lifshitz and Khalatnikov, and to clarify its validity. It may also be that there is some connection between the singularities studied in General Relativity and those studied in other branches of physics (cf. for instance, Thom's theory of elementary catastrophes (1969)). Alternatively one might try to proceed by brute force, integrating the Einstein equations numerically on a computer. However this will probably have to wait for a new generation of computers. One would

like to know also whether the singularities produced by collapse from a non-singular asymptotically flat situation would be naked, i.e. visible from infinity, or whether they would be hidden behind an event horizon.

The other main problem is to formulate a quantum theory of space–time which will be applicable to strong fields. Such a theory might be based on a manifold, or might allow changes of topology. Some preliminary attempts in this line have been made by de Witt (1967), Misner (1969, 1971), Penrose (see Penrose and MacCallum (1972)), Wheeler (1968), and others. However the interpretation of a quantum theory of space-time, and its relation to singularities, are still very obscure.

Speculation and discussion on the subject of this book is not new. Laplace essentially predicted the existence of black holes: 'Other stars have suddenly appeared and then disappeared after having shone for several months with the most brilliant splendour ... All these stars ... do not change their place during their appearance. Therefore there exists, in the immensity of space, opaque bodies as considerable in magnitude, and perhaps equally as numerous as the stars.' (M. Le Marquis de Laplace: 'The system of the world'. Translated by Rev. H. Harte. Dublin, 1830, Vol. 2, p. 335.) As we have seen, our present understanding of the situation is remarkably similar.

The creation of the Universe out of nothing has been argued, indecisively, from early times; see for example Kant's first Antinomy of Pure Reason and comments on it (Smart (1964), pp. 117–23 and 145–59; North (1965), pp. 389–406). The results we have obtained support the idea that the universe began a finite time ago. However the actual point of creation, the singularity, is outside the scope of presently known laws of physics.