THE THEORY OF THE EXPANDING UNIVERSE AS ORIGINATED BY A. A. FRIDMAN

Ya. B. ZEL'DOVICH

Usp. Fiz. Nauk 80, 357-390 (July, 1963)

1. HISTORY OF FRIDMAN'S DISCOVERY OF THE NONSTATIONARY NATURE OF THE UNIVERSE

THIS article is timed for the seventy-fifth anniversary of the birth of A. A. Fridman, which is July 17, 1963.

About 40 years have passed since the publication of the two short papers in which a nonstationary solution of the cosmological problem is given on the basis of the general theory of relativity, or, more simply and briefly, in which the theory of the expanding universe is proposed. Soon, in 1925, Fridman died, before astronomical observations had confirmed his theory. According to the testimony of Academician V. A. Fock, who translated Fridman's work into German for the Zeitschrift für Physik, the author was not sure that the world is actually constructed in accordance with the solutions he had found.*

A correct theory not only explains known facts, but also predicts new phenomena. A classic example which is always given in this connection is the prediction of the existence of the planet Neptune, which was made by the astronomer Leverrier.† But Leverrier used celestial mechanics, which previous to his work had already been brilliantly developed and verified.

On the other hand, Fridman's work was the first (and now we can add, the only) correct application of the general theory of relativity to cosmology.

The general theory of relativity is based on the special theory and the principle of the equality of gravitational and inert masses. In Fridman's theory the natural further assumption is made that the universe is on the average homogeneous and isotropic; in the neighborhood of the solar system and our galaxy other galaxies are distributed with a definite density,* and there is a definite average density of matter; it is assumed that at any distance and in any other place the physical conditions are the same on the average.

Our galaxy is not a specially selected one, central in the universe, just as neither the earth nor the sun is a specially selected body. Furthermore the motions of the galaxies must be such that they do not violate the homogeneity and isotropy of space—that is, the similarity of the conditions at all points of space and the equivalence of all directions in space at a given point.

From this minimum number of assumptions a grandiose conclusion was obtained theoretically: the

*See the preceding paper by V. A. Fock in this issue.

†It is less well known that besides a planet located beyond the orbit of Uranus (i.e., Neptune) Leverrier also predicted the existence of a planet with an orbit inside that of Mercury. Actually there is no such planet. The anomalies in the motion of Mercury from which Leverrier argued are explained by effects of the general theory of relativity.

*Galaxies are concentrations of stars, like the concentration seen as the Milky Way, to which our solar system belongs.

galaxies cannot be at rest relative to each other.* The speed of relative motion of two bodies increases in proportion to the distance between them.

At the present time the maximum speed of recession observed for distant galaxies is 0.3—0.4 times the speed of light—that is, of the order of 100,000 km/sec.

This speed is determined from the Doppler shift of spectral lines.

There is a humorous story of how Robert Wood, after driving through a red light, explained to the policeman that because of the Doppler effect he saw the light as green. In the spectra of distant galaxies, however, the lines corresponding to the blue-green part of the spectrum fall in the red region, and lines are observed in the visible part of the spectrum which under ordinary laboratory conditions are in the ultraviolet.

Considering this effect from the classical point of view, we can say that the kinetic energy of the motion of distant objects is enormously large; this kinetic energy is many times that of any known nuclear reactions (except the reaction of annihilation).

Thus Fridman's theory predicted a grandiose phenomenon, on a scale billions of times larger than that of phenomena in the solar system.

Therefore we can speak without exaggeration of Fridman's scientific achievement as a great one; his work is the basis of all modern cosmology, and its general scientific importance is not less than that of his famous hydrodynamical papers.

The significance and the nontrivial character of Fridman's work are still more evident when one examines the cosmological papers of other scientists, in particular those of the creator of the theory of relativity, Albert Einstein.

Einstein started from the preconceived idea that the universe must be stationary, that is, remain unchanged on the average as time goes on. When it turned out that the equations do not give this kind of solution, he began to make arbitrary changes in the equations of the general theory of relativity (roughly speaking, he introduced something like a negative density and a negative pressure in empty space) for the sole purpose of saving the stationary nature of the solution. Incidentally, Einstein's stationary solution turned out to be illusory—it is unstable with respect to small perturbations.

When Fridman's papers appeared, Einstein's attitude toward them was a negative one. He gave them recognition only after hearing the explanations of Krutkov, who went to Berlin with a letter from Fridman.

*In principle it is possible for the velocity to be zero at a definite time, but the acceleration is always different from zero. Thus the instantaneous condition $v = 0$ is not really a state of rest, just as the vanishing at one point of the velocity of a stone thrown vertically upward is not really a state of rest.

Some years later, in 1929, there followed the experimental confirmation that the universe is nonstationary, and that there is a general recession of the nebulas.

In 1935, in the concluding summary of his book The Meaning of the Theory of Relativity, Einstein pointed out the correctness of Fridman's concept, and emphasized that his own change in the equations was a mistake.

Fridman's papers were published in 1922-1924, in a period of great difficulties. "Russia in a Fog" was the impression H. G. Wells got of Moscow and Petrograd in 1921. In the same issue of the journal as Fridman's article there was an appeal to German scientists to collect scientific literature for their Russian colleagues, who were cut off from the literature during the war and the revolution. Under these conditions the production of a highly significant theory was not only a scientific but also a human achievement.

The present article is addressed to a wide circle of readers not specializing in the field of astronomy. On one hand it includes material which can be regarded as generally known and is contained in textbooks such as Classical Field Theory by Landau and Lifshitz. On the other hand, in a popular article we can dispense with a bibliography, since it is not assumed that all of the views and assertions in the literature are fully expounded.

A detailed bibliography is given in the collection Structure of Stellar Systems (Russian translation, Stroenie zvezdnykh sistem, Moscow, IL, 1963). For papers by the author see the collection Voprosy cosmogonii, Vol. IX (Moscow, AN SSSR, 1963). A number of questions have been treated in detail in an article by Lifshitz and Khalatnikov in the present issue and in an article by the author in this journal [UFN 78, 549 (1962), Soviet Phys. Uspekhi 5, 93 (1963)].

2. THE PAST AND THE FUTURE IN FRIDMAN'S THEORY

As has already been stated, the theory leads to a connection between distance and velocity

$$\mathbf{u} = H\mathbf{r}, \tag{2.1}$$

where H is the so-called Hubble constant, named after the astronomer who discovered the phenomenon of the recession of distant galaxies.

In calling H a constant, one means that H does not depend on the distance between galaxies nor on the direction of $\mathbf{r}$. The theory predicts that H depends on the time. In fact, in inertial motion the velocity is constant, and the distance increases, so that H decreases; besides this there is the effect of gravitation, which also diminishes H.

What does the theory say about the past and the future of the universe?

The conclusions about the past are uniquely determined. At present an expansion is occurring, and con-

sequently the density was larger at earlier times.

The quantity H has the dimensions $\sec^{-1}$. Therefore H^{-1} is a time.

For inertial motion H^{-1} is precisely the time that has passed since the moment when the density was infinite.

When we include the gravitational interaction the speed of expansion in the past is larger than that calculated without gravitation, so that the time T that has elapsed from the moment of infinite density to the present is smaller than H^{-1}: $T < H^{-1}$.

In Fridman's theory the conclusion is unavoidable that there was a time when the density was infinite (it is convenient to take this moment as the origin for reckoning time). This conclusion is valid independent of the form of the law of increase of pressure at large densities; in a homogeneous universe the pressure does not depend on the coordinates, and the only force that affects motion comes from pressure differences (pressure gradient).

The predictions of the Fridman theory about the future depend essentially on the relation between the present values of H (the Hubble constant) and the mean matter density ρ.

There is a definite critical value $\rho_c = (3/8\pi)(H^2/\kappa)$, where κ is the Newtonian gravitational constant

$$\varkappa = 6.7 \cdot 10^{-8} \ \mathrm{cm^3 g^{-1} sec^{-2}}$$

(this expression is derived by means of Newtonian mechanics in the next section).

If the density ρ is actually less than this critical value, $\rho < \rho_c$, then gravitation cannot stop the observed expansion; although the expansion will become slower, it will not be succeeded by a contraction, and the distance between two distant galaxies will increase without limit in the course of time.

If, on the other hand, the density is larger than the critical value, $\rho > \rho_c$ (concerning the actual situation see Section 5), then the attraction is large and the expansion now observed must give way in the future to a stopping and a contraction; instead of a Doppler "red shift" (recession) the astronomers of the distant future will speak of a "blue" or a "violet" shift of spectral lines. In this case the solution gives an infinite density not only in the past, but also in the future.

The time dependences of the distance between two galaxies in the two cases are shown schematically in Fig. 1.

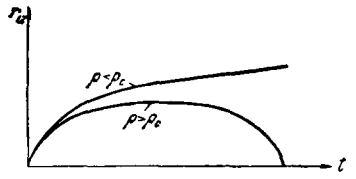

FIG. 1

3. ELEMENTARY DERIVATION OF THE LAW OF EXPANSION

The general theory of relativity contains within it, as a limiting case, classical mechanics together with Newton's theory of gravitation.

Let us consider a small region, inside which all velocities are small in comparison with the speed of light c; the differences of the gravitational potential inside the region are also small (in comparison with c^2). Newton's equations apply to such a region. We shall use a theorem which is equally valid in the Newtonian theory and in the general theory of relativity: matter surrounding a given region in a spherically symmetrical layer has no effect on processes inside the region. It turns out that this statement can be extended to the case of a region conceptually cut out in an infinite space filled with matter with a constant density.

Let us turn to the "arithmetic." We consider a spherical region of radius R, inside which the matter with density ρ has at the time $t = t_0$ velocities distributed according to the law

$$\mathbf{u} = H\mathbf{r}. \tag{3.1}$$

In particular, a particle A at the edge of the region has the instantaneous velocity

$$u_R = HR.$$

The acceleration of this particle is

$$\frac{du_R}{dt} = -\varkappa \frac{M}{R^2}, \tag{3.2}$$

where M is the mass of the matter contained in the region considered,

$$M = \frac{4\pi}{3} \varrho R^3. \tag{3.3}$$

In the course of time ρ and R change, but M is obviously constant. Consequently, it is easy to integrate the equation; we multiply through by $u_R = dR/dt$

$$u_R \frac{du_R}{dt} = \frac{d}{dt}\left(\frac{1}{2} u_R^2\right) = -\varkappa \frac{M}{R^2} \frac{dR}{dt} = \frac{d}{dt} \frac{\varkappa M}{R},$$

$$\frac{1}{2} u_R^2 - \frac{\varkappa M}{R} = \mathrm{const} = k. \tag{3.4}$$

The reader will please pardon the elementary nature of the calculation; we have derived the fact that the sum of the kinetic and potential energies of unit mass at the edge of the region is constant; the important thing is that in the derivation we did not have to use arguments about the work of removing the particle to infinity, which would have depended on what there is around the region.

If at a given time $u_R > 0$ and $k > 0$, then obviously u_R will never become zero, and the expansion will never be supplanted by a contraction.

Let us insert the values of the quantities for $t = t_0$:

$$\frac{1}{2}u_R^2 - \frac{\varkappa M}{R} = \frac{1}{2}H^2R^2 - \varkappa\frac{4\pi}{3}R^3\varrho\frac{1}{R} = R^2\left(\frac{1}{2}H^2 - \frac{4\pi}{3}\varkappa\varrho\right) = k.$$

Thus the critical condition is

$$\frac{1}{2}H^2 - \frac{4\pi}{3}\varkappa\varrho_c = 0, \tag{3.5}$$

$$\varrho_c = \frac{3H^2}{8\pi\varkappa}. \tag{3.6}$$

The result agrees exactly with Fridman's theory. It is also easy to find the laws of time variation of the density and of the Hubble constant both for $\rho > \rho_c$ and for $\rho < \rho_c$.

In order not to engage in much algebra, we confine ourselves to the limiting law for small $t \ll t_0$, for which R is small, $R(t) \ll R(t_0)$, the speed u_R is large, and we can neglect the quantity k in comparison with the large values of u_R^2 and $\kappa M/R$. Then

$$\frac{1}{2}u_R^2 = \frac{\varkappa M}{R}, \quad u_R = \frac{dR}{dt} = \sqrt{\frac{2\varkappa M}{R}}, \quad \frac{2}{3}R^{3/2} = t\sqrt{2\varkappa M}. \tag{3.7}$$

The constant of integration has been chosen so that for $t = 0$ we have $R = 0$, i.e., the origin of time is taken at the instant of infinite density.

Let us find the expression for the density $\rho(t)$. To do so we set

$$M = \frac{4\pi}{3}R^3(t)\,\varrho(t),$$

and obtain

$$\varrho(t) = \frac{1}{6\pi\varkappa t^2} = \frac{8\cdot 10^5}{t^2}. \tag{3.8}$$

The numerical expression is given for density in g-cm^{-3} and t in sec.

Finally, for the Hubble constant we find

$$H = \frac{1}{R}\frac{dR}{dt} = \frac{2}{3}\frac{1}{t}. \tag{3.9}$$

We have considered a region with a quite definite quantity of matter M, with a definite R(t). It has turned out, however, that the results for such quantities as $\rho(t)$ and H(t) do not depend on the choice of M and R. This confirms the internal consistency of the calculation, and the possibility of extending the calculation to infinite space.

There is sometimes talk of a gravitational paradox in the Newtonian theory, of the impossibility of considering an infinite homogeneous universe in this theory. Actually there is a definite sequence of procedures in which no paradox arises. First we shall consider a sphere of finite size R with a definite density ρ and the velocity distribution $u = Hr$; the solution of the corresponding mechanical problem is trivial and leads to definite time dependences H(t) and $\rho(t)$, which do not involve R. Consequently, if we let $R \to \infty$ at a time t_0 with fixed $H(t_0)$ and $\rho(t_0)$, we get a correct solution for an infinite homogeneous universe.

Such a solution could have been obtained only a year after Newton formulated the laws of mechanics and of universal gravitation. Actually this approach was found only in 1935, in the course of thinking through and popularizing the Fridman solution. The Newtonian approach, however, is rigorous and exact.

The solution has been found by considering a spherical region in which a point is singled out—the center of the sphere. At the center O the matter is at rest (Fig. 2). At every other point the matter moves with a definite velocity, and there is a preferred direction —the direction of the velocity u. It is easy, however, to verify on the classical level that this singling out of a center and a direction is fictitious. Let us take an arbitrary point B inside the sphere and go over to a coordinate system in which B is at rest.

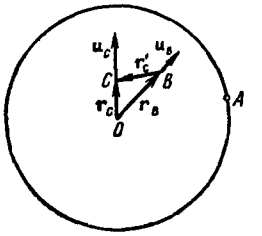

FIG. 2

We distinguish the quantities in the new coordinate system with a prime. It is obvious that for any other point C (see Fig. 2)

$$\mathbf{r}'_C = \mathbf{r}_C - \mathbf{r}_B, \tag{3.10}$$

$$\mathbf{u}'_C = \mathbf{u}_C - \mathbf{u}_B. \tag{3.11}$$

We are using the classical transformation laws: Euclidean for the coordinates, Galilean for the velocities.

We substitute the Hubble law

$$\mathbf{u} = H\mathbf{r},$$

and get

$$\mathbf{u}' = H\mathbf{r}'. \tag{3.12}$$

The law of motion from the point of view of an observer who is at the point B and moves with it does not differ in any way from the law of motion for the observer at the center O of the sphere, with whom we tacitly identified ourselves in the foregoing calculations.

The observer at the point B could say that he is closer to one edge of the region than to the other, but only for the case that the region filled with matter actually has an edge, i.e., is surrounded by empty space. If, however, the region is singled out only conceptually in an infinitely extended homogeneous field of matter density, then the point B is completely equivalent to the center of the sphere, and also to any other arbitrarily chosen point. Thus a solution has actually been constructed which satisfies the principle of homogeneity, but the solution is of necessity a nonstationary one.

It may be that the greatness of Fridman's discovery is not so much that he applied the general theory of relativity as that he broke away from the preconceived idea of the universe as a stationary system.

4. THE CASE OF AN ARBITRARY EQUATION OF STATE

A reader interested only in the general physical aspect of the subject can omit the present section without loss for the subsequent discussion.

Let us consider the changes in the formulas which arise in the case in which the pressure cannot be neglected. In doing so we must compare the pressure with the energy density, in which we must include the rest energy.

If space is filled with particles moving chaotically with the speed of light, for example photons, or neutrinos and antineutrinos, then

$$p = \frac{1}{3}\varepsilon,$$

where p is the pressure and ϵ is the energy density. The quantity which is to be called the mass density is obviously

$$\varrho = \frac{1}{c^2}\varepsilon;$$

consequently for a gas of relativistic particles $p = \rho c^2/3$.

Again we can conduct our treatment in the Newtonian approximation. We must, however, take from the general theory of relativity the fact that the attraction depends on the sum $\rho + (3p/c^2)$ (in the case of a relativistic gas this quantity is equal to 2ρ).

Consequently the acceleration at the radius R is

$$\frac{d^2R}{dt^2} = \frac{du_R}{dt} = -\varkappa\frac{4\pi}{3}\left(\varrho + \frac{3p}{c^2}\right)R. \tag{4.1}$$

A second difference from the previous case is that the mass contained in the sphere of radius R(t), i.e., the mass inside the volume delimited by given particles (in technical language, inside a given Lagrangian radius), does not remain constant during the motion. The point is that when R changes in the presence of a pressure work is done against the pressure forces.

The energy contained in the volume

$$V = \frac{4\pi}{3}R^3,$$

is given by

$$E = \varepsilon V.$$

According to the law of conservation of energy

$$dE = -p\,dV, \tag{4.2}$$

$$\varepsilon\,dV + V\,d\varepsilon = -p\,dV. \tag{4.3}$$

If the pressure is the equilibrium "thermodynamic" pressure

$$p = p(V;\ S) = -\left(\frac{\partial E}{\partial V}\right)_S, \tag{4.4}$$

then as a particular consequence we get the law of conservation of entropy dS = 0.

Again let us integrate the equation of motion; we multiply through by $u_R = dR/dt$:

$$u_R\frac{du_R}{dt} = \frac{d}{dt}\left(\frac{1}{2}u_R^2\right) = -\varkappa\frac{4\pi}{3}\left(\varrho + \frac{3p}{c^2}\right)R\frac{dR}{dt}. \tag{4.5}$$

It turns out that the right member can be written very simply,

$$-\varkappa\frac{4\pi}{3}\left(\varrho + \frac{3p}{c^2}\right)R\frac{dR}{dt} \equiv \frac{d}{dt}\left(\frac{\varkappa}{R}\frac{4\pi}{3}\varrho R^3\right), \tag{4.6}$$

owing to the connection between the change of energy and the pressure. In fact, it is easy to verify that

$$\begin{aligned}\frac{d}{dt}\left(\frac{1}{R}\varrho V\right) &= \frac{1}{c^2}\frac{d}{dt}\left(\frac{1}{R}\varepsilon V\right) = \frac{1}{c^2}\left[\varepsilon V\frac{d}{dt}\frac{1}{R} + \frac{1}{R}\frac{d}{dt}(\varepsilon V)\right] \\ &= -\frac{1}{c^2}\left(\varepsilon V\frac{1}{R^2}\frac{dR}{dt} + \frac{1}{R}p\frac{dV}{dt}\right) = \\ &= -\frac{1}{c^2}\left[\varrho c^2\cdot\frac{4\pi}{3}R^3\frac{1}{R^2}\frac{dR}{dt} + \frac{1}{R}p\cdot 4\pi R^2\frac{dR}{dt}\right],\end{aligned} \tag{4.7}$$

from which Eq. (4.6) follows.

Consequently,

$$\frac{1}{2}u_R^2 - \frac{\varkappa}{R}\frac{4\pi}{3}\varrho R^3 = \frac{1}{2}H^2R^2 - \frac{4\pi}{3}\varkappa\varrho R^2 = \text{const}. \tag{4.8}$$

Since the density is always positive, the question as to whether the velocity can become zero obviously depends on the sign of the constant. Exactly as in the preceding section, we get the same value of the critical density for a given H:

$$\varrho_c = \frac{3H^2}{8\pi\varkappa}. \tag{4.9}$$

It is remarkable that in spite of the change in the expression for the acceleration and in spite of the different law for the variation of the density ($\rho R^3 \neq$ const), the expression (4.8) for "conservation of mechanical energy" has remained unchanged when there is pressure.

Comparison with the expression for the acceleration shows that at the critical density $\rho = \rho_c$ the acceleration in a world filled with radiation ($p = \epsilon/3$) is twice as large as in a world filled with matter with $p \ll \epsilon$.

The expression for "conservation of energy" has been used twice in Sections 3 and 4: for the formula

$$dE = -p\,dV$$

and for the formula

$$\frac{1}{2}u_R^2 - \frac{\varkappa M}{R} = \text{const},$$

that is, as it were, both for the internal energy E and for the mechanical (kinetic and potential) energy. On what grounds can we write two separate conservation equations, and why were the kinetic and potential energies not included in the expression

$$dE = -p\,dV$$

The point is that in a sphere of radius R with the

Hubble velocity distribution the internal energy E is proportional to the volume, i.e., to R^3, and the kinetic energy is proportional to R^5 (since $u = RH$, $u^2 \sim R^2$); in just the same way the gravitational interaction energy $\sim M^2/R \sim R^5$. We are justified in equating separately the terms $\sim R^3$ and the terms $\sim R^5$.

For a given pressure law it is easy to find the concrete solution of the equation of motion.

In the limit of sufficiently high temperature the speeds of all particles will always approach the speed of light, and arbitrarily large numbers of photons and neutrinos are produced, so that actually $p \to \epsilon/3$; therefore it is of physical interest to consider this case.

In this case we have from Eq. (4.3)

$$p \sim \varepsilon \sim \varrho \sim R^{-4}, \quad \varrho = \frac{A}{R^4}, \tag{4.10}$$

where A is a constant.

In Eq. (4.8) we neglect the constant. We get

$$\frac{1}{2}\left(\frac{dR}{dt}\right)^2 - \frac{\varkappa}{R}\frac{4\pi}{3}\frac{A}{R^4}R^3 = 0,$$

$$\frac{dR}{dt} = \sqrt{\frac{8\pi}{3}\frac{\varkappa A}{R^2}}, \quad \frac{R^2}{2} = t\sqrt{\frac{8\pi}{3}\varkappa A}, \quad R = \sqrt{t}\sqrt[4]{\frac{32\pi}{3}\varkappa A}. \tag{4.11}$$

We substitute the value of R in the expression for the density:

$$\varrho = \frac{A}{R^4} = \frac{3}{32\pi\varkappa t^2} = \frac{4\cdot 5\cdot 10^5}{t^2}. \tag{4.12}$$

Finally we have

$$H = \frac{1}{R}\frac{dR}{dt} = \frac{1}{2}\frac{1}{t}. \tag{4.13}$$

As was to be expected, the density and the Hubble constant do not depend on the radius R conceptually chosen for consideration. The expressions for ρ and H do not differ much from the expressions found in Section 3 for $p = 0$.

In a relativistic gas with $p = \epsilon/3$ the speed of sound is $c/3^{1/2}$, from which we can see that this equation of state is not the most rigid one.

The limit of hardness of the equation of state is $p = \epsilon$; in this case the speed of sound is c. This equation of state is obtained for the model of a cold nucleon gas with repulsions between the nucleons owing to neutral vector mesons.

In this case (we give the results without the calculations)

$$p \sim \varepsilon \sim \varrho \sim R^{-6}, \quad R \sim t^{1/3}, \tag{4.14}$$

$$\varrho = \frac{1}{24\pi\varkappa t^2} = \frac{2\cdot 10^5}{t^2}, \quad H = \frac{1}{3t}. \tag{4.15}$$

5. THE STRUCTURE OF THE UNIVERSE AS A WHOLE

In the general theory of relativity it is assumed that physical space is noneuclidean—that it is curved by the presence of matter; the curvature depends on the density and motion of the matter.

It turns out that the critical value of the density on which the future of the universe depends (unlimited expansion or eventual stopping and contraction) is also the critical value for the spatial structure of the universe as a whole.

Our ideas about space depend on the ratio of ρ and ρ_c.

If $\rho < \rho_c$, then space is infinite, and for uniform density the total amount of matter, and thus the number of protons and neutrons, in the universe is also infinite.

If, on the other hand, $\rho > \rho_c$, then space is closed and finite. The explanation of the meaning of a uniform closed three-dimensional space is usually based on the analogy with a closed two-dimensional space. A two-dimensional space is traditionally called a surface; from the point of view of a two-dimensional being the surface of a sphere in three-dimensional space is a closed two-dimensional space.*

Thus if actually $\rho > \rho_c$ the universe is a closed three-dimensional space. Its volume at a given time is finite, and the amount of matter, the number of baryons, in the entire universe has a quite definite value which does not change with time.

On the other hand the volume of the universe changes with time, being proportional to r_{12}^3, the cube of the distance between any pair of distant galaxies (the dependence of r_{12} on t is shown in Fig. 1).

In this sense one speaks not only of the recession (motion away from each other) of the galaxies, but also of the expansion of the universe as a whole. The density of baryons varies as r_{12}^{-3}; if we neglect the pressure, we can say that the density of matter is also proportional to r_{12}^{-3}.

We note that to judge whether the universe is infinite (or, as one says, open) or closed, we compare the present value of the density ρ with the present value of H, on which ρ_c depends.

In the course of time ρ and H change; it turns out, however, that they change in such a way that the sign of the difference $\rho - \rho_c$ cannot change; if it is ever shown that at present $\rho > \rho_c$, this will mean that the same relation has always held and always will hold—the property of being closed or open cannot change in the course of time. Essentially this is evident from the conservation law (4.8), since the sign of $\rho - \rho_c$ is the sign of the constant.

*We note that for a given H the closed or nonclosed nature of the space depends only on the density and not on the pressure, just as in Section 4 the sign of the constants and the character of the future depended only on the density. Ya. A. Smorodinskiĭ has remarked that the physical processes of nuclear reactions, radiation, collisions, and so on can change the pressure, but by the law of conservation of energy do not change the mean density. These processes cannot change the closed or nonclosed nature of the universe.

6. THE DENSITY OF MATTER

The density ρ means the total density of all forms of matter, averaged over all space. A primary contribution is that of the masses of the stars which make up a galaxy, divided by the average volume per galaxy.* Then we must add the mean density of the dust and of the atoms and molecules of hydrogen and other elements in intergalactic space. Finally, ρ also includes the density of such nonclassical types of matter as neutrinos, quanta of the electromagnetic field, and quanta of the gravitational field—gravitons. When the particles or quanta have zero rest mass, the density is the total energy of the particles in 1 cm^3, divided by c^2.

According to the latest data the Hubble constant is about 75 km/sec per megaparsec; these are the most convenient and customary units for astronomers. Translating to cgs units, we have

$$H = 2.5 \cdot 10^{-18}\ \text{sec}^{-1},\quad H^{-1} = 4 \cdot 10^{17}\ \text{sec} = 1.3 \cdot 10^{10}\ \text{years}$$

The corresponding critical density is $\rho_c = 10^{-29}$ g-cm^{-3}.

Estimates of the actual density of matter are based on analyses of the motion of the stars in the galaxies and of individual galaxies in clusters of galaxies.

The speeds of motion are determined from the Doppler effect. There is, however, a definite relation between the kinetic energy of the motion and the potential energy of attraction, which must hold if any system of mutually attracting bodies is to have a prolonged existence (the virial theorem). Thus we can determine the total mass in the volume of the system. This gives an actual determination of the sum of all kinds of matter in the volume of a galaxy.

We assume that the greater part of the mass is concentrated in the stars. Then the intergalactic space, in which there are almost no stars, is to be regarded as practically empty.

An empirical relation can be established between the total luminosities and the masses of galaxies. By taking the ratio of the masses of the galaxies to the total volume, Oort arrives at a probable value for the mean density of about $\rho \sim 3 \times 10^{-31}$ g-cm^{-3}, which is less than ρ_c by a factor 30.

This would mean that the universe is infinite and that the expansion observed at the present time will never cease. Astronomy has, however, seen repeated revisions of quantitative estimates of this sort. Therefore Oort's result cannot be regarded as conclusive.

7. OBSERVATIONAL INVESTIGATION OF THE STRUCTURE. THE HORIZON

Another approach to the investigation of the universe is to study distant nebulas in terms of the amounts of light and radio-frequency radiation which come to us, the angles at which these objects are seen, and the magnitude of the red shift. The main purpose is to determine the curvature of space; on certain assumptions about the galaxies, their sizes and luminosities, the observed distributions will depend on the relations between the surface area and the volume of a sphere and its radius which hold in the space of the universe.

We note that owing to the nonstationary nature of the universe the two very different types—the closed and the infinite worlds—do not lead to qualitative differences in the observed pattern.

The farther from us an observed object is located, the longer the time that light has required to reach the observer. The light observed today was emitted at an earlier time. But the entire evolution of the universe from the instant of infinite density to the present day has taken a definite time, not more than 10^{10} years. Therefore even in the infinite world there is a "horizon"—a limiting distance, corresponding to a time of propagation of light of about 10^{10} years, beyond which are located objects which in principle cannot be observed today; as time goes on, this horizon broadens by about 10^{-6} percent in 100 years.

Inside, in the region accessible to observation, there is a definite, finite amount of matter, stars, galaxies. As the horizon is approached the red shift becomes stronger, so that the observed frequency of the light approaches zero. There comes to mind a poetic comparison with the visible reddening of the sun at sunset, on the horizon, but the actual cause is different in the case of the sun!

Thus in passing the Fridman theory solved a paradox which had excited astronomers for more than 100 years: if the universe is infinite, it would seem that in any element of solid angle we must—at some distance, smaller or larger—see the surface of some star. Then the entire sky would have a brightness of the order of that of the sun! Actually the density of stars is small, and almost everywhere we see a prestar condition of matter of large density. Even if the matter were hot, the red shift would be enough to cool down the photons, decrease their energy, and render them invisible.

Let us pass on to the case of a closed world. In a stationary closed world it would be possible to see the same distant object twice.

We turn to our two-dimensional analogy: light propagates on the surface of a sphere along the shortest, so-called geodesic, lines.

If we are at the north pole, the family of geodesic lines—rays—comprises the meridians. Thus a given object A can be seen both by means of the ray emitted from A toward the north, and also by means of the ray that goes from A through the south pole; the second ray comes to the observer at the north pole from the opposite direction.

We know as a fact, however, that the universe is nonstationary and that at present we are in a stage

*Not by the volume of the galaxy itself, but by a volume that depends on the mean distance between galaxies!

of expansion. It then turns out—in the case of a closed world—that the time that has passed from the instant of infinite density to the present is smaller than the time for propagation of light from one pole to the other.

Thus if the world is closed, at the present time our horizon takes in only part of the entire (finite) mass of the world; the possibility of a double observation of a single object is excluded a fortiori. Thus both in the case of the open world and in that of the closed world there is in principle a threshold of observation, a "horizon," corresponding to zero frequency of the observed light and infinite density of the matter emitting the light.

For a given value of the Hubble constant, depending on the value of the density, the homogeneous world is infinite for $\rho < \rho_c$ and closed for $\rho > \rho_c$; as ρ approaches ρ_c from above (through values $\rho > \rho_c$) the radius of the world and the quantity of matter contained in it go to infinity, and the value $\rho = \rho_c$ is in fact a critical one for the most important properties of the world as a whole. For all quantities, however, that can in principle be observed at the present time (for example, for the amount of matter in the sphere of observation, up to the horizon), the value $\rho = \rho_c$ is not in any way singled out as a special value—there are no qualitative changes that depend on whether $\rho > \rho_c$ or $\rho < \rho_c$.

This does not exclude a quantitative, functional dependence of observable quantities on the density ρ. The relation between the red shift and the distance depends on the value of ρ. The expression

$$\frac{\Delta\omega}{\omega} = \frac{u}{c} = \frac{Hr}{c} \qquad (7.1)$$

(ω is the frequency, $\Delta\omega$ is the change of frequency) is only a first approximation, the first term of an expansion of $\Delta\omega$ in terms of the distance.

In a curvilinear space with a nonstationary (time-dependent) metric the very definition of distance is not unambiguous.

Therefore what must actually be done is to find, in terms of the Hubble constant and the density, the relations between quantities which can (at least in principle) be uniquely determined from observations. Quantities of this sort are:

1) the change of frequency of spectral lines emitted by an object;

2) the angular diameter at the point of observation of an object whose absolute linear size is regarded as known;

3) the visible brightness (flux of luminous energy perceived by the observer) of an object whose absolute luminosity is regarded as known;

4) The total amount of matter inside a sphere drawn through a given object (with its center at the point of observation).

The predictions about angular diameters are particularly curious: at small distances, in the Euclidean limit, we obviously have

$$\theta = \frac{l}{r} = \frac{lH}{c\frac{\Delta\omega}{\omega}}, \qquad (7.2)$$

where θ is the angle, l is the absolute size, and r is the distance, which can be expressed in terms of $\Delta\omega$ and H; naturally θ decreases with increasing red shift. As the horizon is approached, however, r decreases, since the horizon takes in a finite amount of matter and the density of matter goes to infinity as the horizon is approached. Therefore θ goes through a minimum and increases without limit for $\Delta\omega/\omega \to 1$ (see note added in proof at end of article). This result also is the same for both the closed and the infinite worlds. Consequently the comparison with observations is based on definite assumptions about the observed objects; for example, for variable stars (cepheids) it is assumed that the absolute luminosity bears a definite relation to a readily observable quantity, the period of variation of the brightness.

Cepheids, however, can be observed only at small distances. For large distances it is necessary to make definite assumptions about the absolute luminosities and sizes of entire galaxies of particular types.

By means of optical observations Baum in 1957 obtained a result which corresponds to a matter density of $\rho = (2 \pm 1)\rho_c$; if we assume that ordinary matter makes up a small part of the density and that the greater part of it comes from neutrinos and photons, we get from these same observational data the result $\rho = (1 \pm 0.5)\rho_c$. These data tend to favor a closed world, and in any case they indicate that the total density is much larger than that calculated from the galaxies—that is, they suggest that there is a considerable amount of matter in intergalactic space. It must be kept in mind, however, that very recently (Sandage) there has been a tendency to lower the value of the density calculated in this way to $(0.2—0.1)\rho_c$ (because of the evolutionary effect, see below).

Radio telescopes are regarded as extremely promising for research on the structure of the universe. Their sensitivity is so great that one can detect and locate on the celestial sphere radio galaxies that are not visible with the most powerful optical telescopes.

Calculations indicate that the most powerful radio galaxies now known could be observed even if they were at a distance at which the red shift reduces the frequency to 0.25 of the emitted frequency. At the horizon the frequency is reduced to 0; consequently, radio telescopes make it possible to detect powerful radio galaxies in a volume which is much more than half of the total volume observable in principle.

For radio galaxies, however, the frequency spectrum is continuous; therefore the red shift cannot be directly measured. There remains the statistical study of the distribution of radio galaxies with respect to their visible brightnesses.

In Euclidean space in a stationary universe the visible brightness is given by $I = L/R^2$, and the volume of a sphere is $4\pi R^3/3$. We assume that L is the same for all galaxies.

The galaxies with brightnesses larger than I are those at distances $R \leq (L/I)^{1/2}$, and their number is proportional to $R^3 \sim L^{3/2}I^{-3/2}$, so that $N(I) \sim I^{-3/2}$.

The nonstationary nature and the curvature (non-euclidean nature) of the universe decidedly alter this distribution law N(I).

For a given value of H the law depends on the density ρ. When we calculate for a small visible brightness, i.e., for $I \to 0$, N(I) increases more slowly than by the classical law $I^{-3/2}$.

Actually the observations show that N(I) increases faster then $I^{3/2}$ for $I \to 0$.

The cause of this deviation can in principle be understood: if the universe as a whole is nonstationary there are no grounds for supposing that the properties and numbers of radio galaxies remain unchanged on the average. For the study of the structure of the universe we need observations of objects at the greatest possible distances, that is, of objects in early stages of their history.

Consequently observation shows that in the remote past the conditions for radio emission were more favorable than at present; radio galaxies were a larger percentage of the total number, and were brighter on the average.

On the other hand, it can be seen from this that without a concrete theory of the radio emission the function N(I) cannot be used for the study of the curvature of space; the evolutional effect is large.

It was mentioned above that for the optical observations, where the influence of the evolutional effect is smaller than in radio astronomy, including this effect can lead to a reduction of the calculated density by a factor of 5 to 10.

8. NEUTRINOS AND ELECTROMAGNETIC RADIATION

How do matters stand with the direct determination of the densities of the various forms of matter?

In recent years there is much interest in the problem of the cosmological density of neutrinos.

Nuclear reactions in stars involve the emission of energetic neutrinos, which can in principle be detected by nuclear methods; their detection is the problem of neutrino astronomy—a new science, and experimentally a very difficult one. At least in principle, neutrino astronomy gives a unique possibility for testing whether or not galaxies and stars of antimatter exist in the universe. The light quanta which they emit do not differ from those emitted by matter, and cannot be used to distinguish antimatter. As we know, however, neutrinos and antineutrinos are different, and cause different reactions under terrestrial laboratory conditions. Therefore antistars can be distinguished from stars!

But we have strayed from our main theme—the density of matter in the universe.

The neutrino emission from nuclear reactions is on the whole of the same order of magnitude as the total energy release in the reactions, and consequently of the same order as the light emission from the stars. It is known that the intergalactic density of starlight is small, smaller than the mean density of matter by at least a factor 1000.

We can get another estimate by starting from the fact that the total energy release in the combination of all nuclear reactions and gravitational contraction must amount to a small fraction of the rest energy of the nucleons of which the stars are composed. Furthermore, at present evidently more than half of the matter is still in the form of hydrogen, and has not yet been involved in nuclear reactions. From this we can draw the very general conclusion that the energy density in starlight and neutrinos from stars is much smaller than the density contributed by the stars.

The situation is different with the low-energy neutrinos. If the world were hot (we shall speak of this later) and if there were a large density of neutrinos and antineutrinos in thermal equilibrium, then in the course of the expansion the energy and the temperature would decrease according to the law of adiabatic expansion. It can be assumed that at present there are neutrinos and antineutrinos with an effective temperature amounting to 20°K. This is enough so that the density of neutrinos would be 10 times the probable matter density of 3×10^{-31} g-cm^{-3}. Consequently, such soft neutrinos would have a very strong effect on the structure of the universe through their gravitational action, but by the methods of nuclear and atomic physics it is entirely impossible either to detect them or to show that they do not exist.

Even the density of energy in such a trivial form as electromagnetic radiation has not been completely investigated.

The radiofrequency region of wavelengths larger than a few centimeters has been well studied. On the other hand, the radiation in the optical region is also well known. At both ends of the spectrum the intergalactic energy density (divided by c^2) is extremely small in comparison with the density of matter.

It is now really necessary to complete the study of the electromagnetic spectrum: the intensity must also be measured in the range of wavelengths from 1 cm to 0.01 cm, so as to rule out the possibility that the radiation density in the cosmos is appreciable.

In this way we shall also get indirect information about the density of neutrinos: when the density is high an equilibrium becomes established, and when the temperature is the same the energy densities of the light and the neutrinos must be in the constant ratio 1:7/4. Both densities will then decrease in the course of the

expansion according to the same law, and the ratio between them stays the same, regardless of what the mechanism for maintaining the equilibrium may be.

Let us assume that thermal radiation will be found with a Planck spectrum corresponding to temperature 15°K (see note added in proof at end of article). Its energy maximum is at a wavelength of about 2.5 mm.

Then there is reason to suppose that the density and mean energy of the neutrinos and antineutrinos is of the same order as the density and mean energy of the photons (see Section 12)—of the order of 10^{-30} g-cm^{-3}, that is, 10^{-9} erg-cm^{-3}.

9. THE METRIC OF THE EXPANDING UNIVERSE

In the general theory of relativity the Fridman solution is characterized by the following expression for the interval:

$$ds^2 = c^2 dt^2 - a^2(t)\,[dr^2 + f^2(r)\,(\sin^2\theta\, d\varphi^2 + d\theta^2)]. \quad (9.1)$$

In this formula a particle moving with the mean Fridman (Hubble) velocity is characterized by constant values of the coordinates r, θ, φ. These coordinates are called the comoving system, and correspond to the Lagrangian system of coordinates in hydrodynamics.

The function f(r) characterizes the curvature of space and depends on the density:

$$\left.\begin{aligned} f(r) &= \operatorname{sh} r, & \varrho &< \varrho_c, \\ f(r) &= r, & \varrho &= \varrho_c, \\ f(r) &= \sin r, & \varrho &> \varrho_c. \end{aligned}\right\} \quad (9.2)^*$$

It must be noted at once that in all cases for small r we have f(r) = r. The behavior of the function a(t) also depends on the density.

Locally the interval is given by the expression

$$ds^2 = c^2 dt^2 - dx^2 - dy^2 - dz^2, \quad (9.3)$$

or in a spherical coordinate system

$$ds^2 = c^2 dt^2 - dR^2 - R^2(\sin^2\theta\, d\varphi^2 + d\theta^2). \quad (9.4)$$

This means that in the expression (9.1) t is the time measured in the comoving system. The distance between particles in physical space is proportional to a(t); for example, the distance from the origin is

$$R = a(t)\, r. \quad (9.5)$$

For a particle at rest in the comoving system, r = const; this means that its velocity is

$$u = r\frac{da}{dt} = \frac{R}{a}\frac{da}{dt}, \quad (9.6)$$

from which it follows that the Hubble constant can be expressed in terms of a:

$$H = \frac{1}{a}\frac{da}{dt}. \quad (9.7)$$

That the space is curved is indicated by the fact that f(r) is different from r at large distances. The length of the circumference is given by

$$L = 2\pi a(t)\, f(r),$$

whereas the radius of the circle is R = a(t)r. Obviously L can be different from $2\pi R$ only in a noneuclidean space. It must be emphasized that the discontinuous change in the form of f(r) for $\rho < \rho_c$, $\rho = \rho_c$, $\rho > \rho_c$ does not lead to any jump in the metric of the neighborhood, since for $\rho \to \rho_c$ we have $a(t) \to \infty$.

Indeed, let us find the circumference of a circle of radius R, using only two terms in the expansion:

$$\operatorname{sh} r = r + \frac{r^3}{6}, \quad \sin r = r - \frac{r^3}{6}, \quad R = ar,$$

$$L = 2\pi a\left(r \pm \frac{r^3}{6}\right) = 2\pi\left(R \pm \frac{R^3}{6a^2}\right). \quad (9.8)$$

The sign + or − depends on $\rho < \rho_c$ or $\rho > \rho_c$; for $\rho = \rho_c$ we have $L = 2\pi R$, and the world is flat.

Since, however, $a \to \infty$ and $1/a^2 \to 0$ for $\rho \to \rho_c$, the transition is actually a smooth one.

10. CURVATURE OF SPACE, SUPERLIGHT SPEED, AND THE MILNE MODEL

The use of the proper time of the co-moving coordinate system leads to certain paradoxes.

It is well known that in the general theory of relativity the curvature of space depends on the presence of matter, which produces a gravitational field.

Then why is the space flat for a definite value of the density, $\rho = \rho_c$, whereas when the density is zero, $\rho = 0$, we can show that a(t) = ct but the metric is hyperbolic, f(r) = sinh r? The whole point here is the choice of the time.

Let us analyze in detail the case with $\rho = 0$.

We consider a conceptual experiment: at a definite instant particles fly out from a definite point with all possible velocities (including velocities smaller than but arbitrarily close to c). We neglect the mass of these particles and the gravitational field that they produce. Consequently, the particles do not change the metric of the space and themselves move along straight lines with constant speeds.

The trajectories of these particles are shown in Fig. 3. The abscissa is the coordinate Z (or r, with fixed θ, φ along a given ray), and the ordinate is the common laboratory time τ.

We shall now consider this same motion in a coordinate system which accompanies the motion of the particles. In this system the time t is the time measured by a clock which moves along with a particle. By the well known transformation law for time in the special theory of relativity

$$t = \tau\sqrt{1-\beta^2} \quad (10.1)$$

("time flows more slowly in a moving system"). Here

$$\beta = \frac{V}{c} = \frac{r}{\tau c}. \quad (10.2)$$

*sh = sinh.

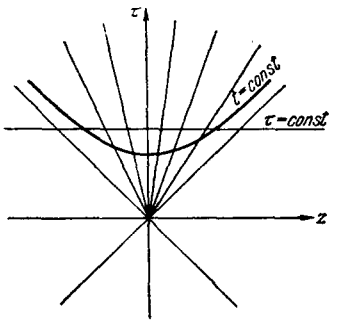

FIG. 3

At the edges of the light cone $t \to 0$. We draw on the diagram a line t = const. This line is a hyperbola and goes out to infinity. It is obvious that for $\rho = 0$ the four-dimensional space-time is flat, the tensor $R_{iklm} = 0$.

The curvature of the three-dimensional space, however—of the hypersurface orthogonal to the fourth coordinate, the time—depends on how the time is chosen. When the "laboratory" time τ is chosen the three-dimensional space is flat, but when we choose the time t of the comoving system the three-dimensional space corresponding (orthogonal) to it is a curved, hyperbolic, space.

It can also be seen from this example what is the meaning of the frightening infinite velocity, velocity larger than c, in the Fridman solution. In fact, if we define the velocity, for example, as $u = (d/dt)(L/2\pi)$, where L is the circumference* and t is the proper time, it is not surprising that $u > c$. Owing to the contraction of time a particle can traverse any path in unit proper time, with a laboratory velocity approaching the speed of light.

Precisely this was observed 15 years later in the study of fast μ mesons in cosmic rays. With a lifetime of $2 \cdot 10^{-6}$ sec they traversed without decay a path much larger than $2 \cdot 10^{-6} \cdot 3 \cdot 10^{10} = 6 \cdot 10^4$ cm = 600 m.

In the literature the case $\rho = 0$ carries the special name of the Milne model.

As we see, the concepts of curved or flat three-dimensional space are to a considerable extent conventional, depending on the way the time is chosen.

By choosing a time like the "laboratory" † time in the case $\rho = \rho_c$, we reach the conclusion that space is not at all flat.

This conventional character, however, attaches only to the noninvariant concept of three-dimensional curvature. The four-dimensional curvature is not a conventional concept; it is zero for $\rho = 0$ and different from zero for $\rho = \rho_c$ in the so-called "flat Fridman world."

*$L/2\pi$ is a definition of the radius which is the same in the laboratory and comoving systems.

†The so-called Schwarzschild time, orthogonal to R = a(t)r, and not to r.

There is also a quite definite meaning of such a quantity as the total number of baryons in the universe, in cases in which we take the point of view of complete homogeneity of the physical conditions and the absence of boundaries, even beyond the horizon.

If $\rho = \rho_c$, so that the world is closed, the number of baryons is equal to their density at a given instant in our neighborhood, n(t), multiplied by the volume of the closed world, as calculated from the instantaneous value of the radius* a(t):

$$N = nV = n \cdot 2\pi^2 a^3(t). \qquad (10.3)$$

To calculate N one uses the values for a particular time t, but the quantity N does not depend on t and does not depend on the way the time and the three-dimensional section are chosen.

For $\rho < \rho_c$, which corresponds to an open world, $N = \infty$, and this too is an objective fact, which does not depend on the way the section is chosen.

11. RED SHIFT, OR AGING OF QUANTA?

The main observational proof of the nonstationary character (expansion) of the universe is contained in the red shift of spectral lines emitted by distant objects. The lowering of the temperature and the decrease of the energies of photons and neutrinos, which has been mentioned, is also a special case of the "red shift." Therefore it is useful to give the red shift law more detailed consideration in all of its aspects.

For the same reason—that the red shift is so important in principle—it has been severely subjected to attacks and doubts.

Cannot we find some different explanation of the red shift, which would allow us to escape the conclusion that the universe is nonstationary?

From time to time more or less vague ideas are put forward about the "aging" of photons, about some sort of mechanism of energy loss from the photons, in which the fraction of the energy lost increases with the distance travelled by the photon. There are at least three very weighty objections to such ideas:

1) If the energy loss occurs owing to an interaction with the intergalactic matter, it is accompanied by a transfer of momentum; that is, there is a change of the direction of motion of the photon.

There would then be a smearing out of images; a distant star would be seen as a disk, not a point, and this is not what is observed.

2) Let us suppose that a photon decays, $\gamma = \gamma' + k$, giving up a small part of its energy to some sort of particle k. It follows from the conservation laws that k must move in the direction of the photon (this, by the way, avoids a smearing out) and have zero rest mass. Because of the statistical nature of the process, however, some photons would lose more energy than

*The radius a(t) can be expressed in terms of the Hubble constant, the density, and the pressure.

others, and there would be a broadening of the lines, which is also not observed.

3) Finally, the most important theoretical argument is due to the Leningrad physicist M. P. Bronshteĭn, who was lost to us by an untimely death.

We ask the question: if there were such a process, how could the decay probability w for a photon depend on its frequency?

At first glance we would have from the dimensions $w = A\omega$, since w and ω have the same dimensions (sec^{-1}): a definite probability of decay per vibration.

Actually, as Bronshteĭn showed, the only possible answer is $w = B/\omega$, with a dimensional constant B (sec^{-2}). The point is that the individual vibrations in a light wave are perceived only by an observer past whom the wave travels. The frequency is different for a moving observer. Bronshteĭn's result follows from the special theory of relativity. It is most simply derived by thinking of the well known relation between the lifetime and the energy of a particle, as verified experimentally for mesons (μ and π).

It is well known that $T = T_0(1-\beta^2)^{-1/2}$, where T is the lifetime of a moving meson as measured by a stationary observer, T_0 is the lifetime of a stationary meson, and $\beta = v/c$, where v is the speed of motion and c is the speed of light. On the other hand the energy of the moving meson is

$$E = \frac{m_0 c^2}{\sqrt{1-\beta^2}}, \tag{11.1}$$

where m_0 is its rest mass. Consequently, we can relate the decay probability w to the energy of the moving meson:

$$w = \frac{1}{T} = \frac{\sqrt{1-\beta^2}}{T_0} = \frac{m_0 c^2}{T_0 E} = \frac{A}{E}. \tag{11.2}$$

This relation is a universal one which follows from the Lorentz transformation. For the photon we must assume that $m_0 \to 0$, but simultaneously $T_0 \to 0$, so that the ratio m_0c^2/T_0 has a definite value. In a particular coordinate system we express the energy of the photon in terms of its frequency

$$E = \hbar\omega,$$

and we get

$$w = \frac{A}{\hbar\omega} = \frac{B}{\omega}.$$

Thus if the decay of photons is possible at all, those in radio waves must decay especially rapidly!

This would mean that the Maxwell equations for a static electric field would have to be changed (since this is the limit of radio waves as the frequency goes to zero), and this is very unpleasant.

There is no experimental indication of such effects: the radio-frequency radiation from distant sources is transmitted to us not a bit more poorly than visible light, and the red shift measured in different parts of the spectrum is exactly the same— $\Delta\omega/\omega$ is constant and corresponds to a single velocity.

Thus suggestions that there is an explanation of the red shift other than Fridman's fail completely.

12. THE RED SHIFT. EXACT FORMULA AND PHYSICAL INTERPRETATION

Let us consider the propagation of a light ray from one star with the coordinates r_1, θ, φ to another star with the coordinates r_2, θ, φ. The choice of equal values of θ, φ does not affect the result. It is important that a star that takes part in the general Hubble motion and does not have any "peculiar" velocity relative to neighboring galaxies will have constant values of r, θ, φ, since these are the comoving Lagrangian coordinates.

The fundamental equation for the propagation of the light ray is ($d\theta = d\varphi = 0$)

$$ds^2 = 0 = -c^2 dt^2 + a^2(t)\, dr^2, \qquad r_2 - r_1 = \int_{t_1}^{t_2} \frac{dt}{a(t)}. \tag{12.1}$$

The notations are shown in Fig. 4.

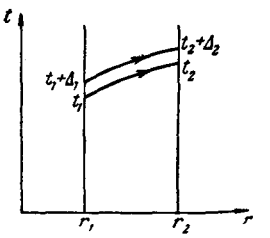

FIG. 4

The trajectory of the star emitting the light is the left-hand vertical line; that of the star receiving the light is the right-hand line; and the trajectory of the light ray is a sloping line with an arrow.

Let us now consider a second ray, emitted later, after the time Δ_1. It also arrives later, by the time Δ_2. From Eq. (12.1) we obviously have

$$\int_{t_1+\Delta_1}^{t_2+\Delta_2} \frac{dt}{a(t)} = \int_{t_1}^{t_2} \frac{dt}{a(t)}, \tag{12.2}$$

from which it follows that (Δ_1 and Δ_2 are small)

$$\frac{\Delta_1}{a(t_1)} = \frac{\Delta_2}{a(t_2)}. \tag{12.3}$$

Suppose that during the time Δ_1 there have occurred a definite number of vibrations of the electron emitting the light. This same number of vibrations of the electromagnetic field is registered by the receiver. The vibration frequency ω is inversely proportional to the time required for a definite number of vibrations. Consequently

$$\frac{\omega_2}{\omega_1} = \frac{a(t_1)}{a(t_2)}. \tag{12.4}$$

We recall that a is the radius of the world, and all distances are proportional to a. In the expanding

world a increases with the time, and $a(t_2)$ is larger than $a(t_1)$; consequently, the frequency received is smaller than that emitted, and we have a red shift.

Let us consider the stars more closely and show that the formula is that of the Doppler treatment.

Let the distance be

$$R = a(r_2 - r_1);$$

$$t_2 = t_1 + \frac{R}{c}, \quad \frac{R}{c} \ll t,$$

$$\frac{\omega_2}{\omega_1} = \frac{a(t_1)}{a\left(t_1 + \frac{R}{c}\right)} = \frac{a}{a + \frac{R}{c}\frac{da}{dt}} = 1 - \frac{R}{c}\frac{1}{a}\frac{da}{dt}. \qquad (12.5)$$

But we saw that the Hubble constant is given by

$$H = \frac{1}{a}\frac{da}{dt}, \quad u = HR. \qquad (12.6)$$

Consequently

$$\frac{\omega_2}{\omega_1} = 1 - \frac{R}{c}H = 1 - \frac{u}{c}. \qquad (12.7)$$

But this is precisely the Doppler formula for the frequency shift for

$$u \ll c.$$

Recently there has been experimental confirmation of one of the predictions of the general theory of relativity: the change of frequency of photons as they move in the gravitational field of the earth has been measured.

Is this effect included in the expression for the red shift? Of course it is included, since Fridman's theory and his calculation of the red shift were developed consistently on the basis of the general theory of relativity.

Let us consider a sphere of radius R with a receiver at the center and a source of light at the surface of the sphere. From the point of view of an observer at the center the energy of a photon increases as the light falls from the surface to the center.

If the distance R is small, the gravitational effect increasing the frequency of the photon is proportional to R^2; the potential difference between the center and surface of the sphere is $-\frac{1}{2}\kappa M/R$, and

$$M = \frac{4\pi}{3}\varrho R^3,$$

from which it follows that the effect is proportional to R^2.

Meanwhile the Hubble velocity and the Doppler effect are proportional to R. Therefore for small R we are correct in speaking of a Doppler effect.

The formula (12.4) is entirely exact, and includes all effects; the matter density ρ, on which the gravitational effect depends, affects $a(t)$ (as we have seen, the "acceleration" d^2a/dt^2 depends on ρ).

It is easy to verify qualitatively that the expression (12.4) contains the gravitational effect. Let us suppose that $\rho > \rho_c$, and consider the emission of light at the time when a is a maximum,

$$t_1 = t_m, \qquad a(t_1) = a(t_m) = a_m. \qquad (12.8)$$

At this instant H = 0, all velocities are zero, and there is no Doppler effect.

At the time when the light is received, however,

$$a(t_2) < a_m,$$

$$a(t_2) = a\left(t_2 + \frac{R}{c}\right) = a(t_m) + \frac{1}{2}\left(\frac{R}{c}\right)^2 \frac{d^2a}{dt^2} < a(t_m)$$

$$\left(\frac{da}{dt} = 0 \quad \text{for} \quad t = t_m, \quad \frac{d^2a}{dt^2} < 0\right); \qquad (12.9)$$

so that

$$\omega_2 > \omega_1, \qquad (12.10)$$

and when there is no Doppler effect there remains only the gravitational blue shift proportional to R^2.

The expression for the red shift can also be interpreted in a different way.

Let us think of a closed world, in which there is a standing electromagnetic wave of a certain n-th order mode, in which, for example, there are n nodal surfaces where the field is zero; we regard this world as a cavity resonator and study a definite harmonic in it. it.

The wavelength λ is a definite fraction of the radius a of the world, $\lambda = a/n$, and the speed of light is a world constant; therefore

$$\omega = \frac{2\pi}{\lambda} = \frac{2\pi n}{a} = \frac{\text{const}}{a}.$$

It is obvious that as the world expands there is no change in the mode and the order n, and we conclude that

$$\omega \sim \frac{1}{a}.$$

The "horizon" corresponds to the emission of the light at the "beginning," i.e., the instant of infinite density and zero radius; it follows from the red-shift formula that light emitted with a finite frequency at the horizon will be received by us today with zero frequency.

Finally, one more aspect of the "red shift" opens up if we consider an assembly of photons which are in equilibrium with a definite temperature T. As is well known, in such an assembly the energy density is given by

$$\varepsilon = 7.6 \cdot 10^{-15} T^4 \text{ erg cm}^{-3}$$

(T in degrees); the (Wien) maximum occurs for photons of frequency

$$\omega = 5 \cdot 10^{11} T \text{ sec}^{-1}.$$

According to the fundamental formula, during the expansion the frequency decreases in inverse proportion to the radius $a(t)$.

It is remarkable that during the course of the red shift and the expansion the Planck distribution remains an equilibrium distribution. There is just a similar decrease of the temperature, as $[a(t)]^{-1}$. For the energy density we have $\epsilon \sim T^4 \sim a^{-4}$. But

the volume of the world* is $V \sim [a(t)]^3$. Accordingly,

$$\varepsilon = \text{const} \cdot V^{-4/3}. \qquad (12.11)$$

This is nothing other than the well known law of adiabatic expansion of a gas with $p = \epsilon/3$, i.e., with the adiabatic index $\gamma = 4/3$:

$$p = (\gamma - 1)\,\varepsilon.$$

Thus the adiabatic cooling in the course of the expansion occurs without any interaction of the photons with dust, atoms, or electrons.

The law of frequency change, which can be derived kinematically (from classical ideas about the Doppler effect and the Hubble expansion) or from the general theory of relativity, leads to the thermodynamic result.

This is also precisely the way the change of energy occurs for an assembly of neutrinos and antineutrinos that are in thermal equilibrium, if they do not interact with anything.

Let us assume that the photons and neutrinos were in thermodynamic equilibrium and had the same temperature at some instant, and that during the expansion they had ceased to interact with other particles and with each other.

According to the expansion law $T \sim a^{-1} \sim V^{-1/3}$ and the temperature ratio T_γ/T_ν always remains equal to 1. The ratio of the energy densities $\epsilon_\gamma/\epsilon_\nu = 0.57$ will not change with time.†

13. ELECTRIC CHARGE IN THE UNIVERSE

The electric charge of a closed world is identically zero. This assertion sometimes raises doubts: what if one electron were added?

The point is that in a closed world there does not exist for an uncompensated charge any solution of the Poisson equation $\operatorname{div} \mathbf{E} = \Delta\varphi = 4\pi\rho_e$, which connects the electric field $\mathbf{E}$, the potential φ, and the charge density ρ_e, with $\int \rho_e\, dV$ not equal to zero.

Speaking intuitively, according to Gauss's theorem the charge is the source of lines of force of the field; emerging from a charge, these lines either go to another charge of the opposite sign, or run off to infinity. In a closed world, however, there is no infinity, and consequently it is inevitable that charges of different signs must compensate each other.

In the homogeneous open infinite world there is an analogous, though weaker, theorem; the mean charge density per unit volume (or per baryon) is zero. Thus it is not excluded that there is a finite total charge, but an infinite total electric charge is impossible. In fact, if at some point the electric field is zero but the mean charge density is not zero, then as we go away from that point the field increases without bound. This contradicts the homogeneity of the world (the field is different in value at different points), and also its isotropy: the electric field is a vector, and marks a definite direction in space.

The similarity between electrostatic and gravitational interactions is emphasized in elementary physics. Why is a constant density of electric charge impossible and a constant density of gravitating matter possible?

The point is that the gravitational field (in the narrow sense of a force vector acting on a unit mass) is not an objectively existing invariant quantity. According to the principle of equivalence, by going over to a coordinate system moving with an acceleration we can locally free ourselves from any gravitational field. It is just with this acceleration that the particles move relative to each other in the Fridman solution; therefore at pleasure we can declare any point to be at rest, and set the gravitational field equal to zero at any point. This cannot be done with an electric field—the difference between the forces acting on an electron and on a proton does not disappear in any coordinate system.

14. ASYMMETRY WITH RESPECT TO CHARGE

As we have seen in the preceding section, in both cases, the open and the closed, the world is electrically neutral. Obviously, however, the world is not charge-symmetrical—the negative charges are mainly electrons, but the positive charges are protons, not positrons.*

It may suitably be emphasized that the fact that the world is not charge-symmetrical does not contradict the complete charge symmetry of the properties of the individual particles. The lack of symmetry is possible as a consequence of the presence of baryons and absence of antibaryons in the initial state.

In mathematical language we have to do with the

*In the case of an open world the total volume is infinite. We could then speak of the volume of a definite region, bounded by given nebulas – this volume is again proportional to a^3.

†We note that because of the conservation of leptons the equilibrium of neutrinos and antineutrinos is determined not only by the temperature, but by what we may call the leptonic charge or the lepton number

$$q = n_\nu - n_{\tilde{\nu}} \quad (q, n\text{—in particles per cm}^3),$$

$$\varepsilon_\nu = \text{const} \cdot T^4 f\left(\frac{|q|}{T^3}\right).$$

During the expansion

$$q \sim a^{-3} \sim V^{-1}, \quad \frac{|q|}{T^3} = \text{const},$$

so that

$$\varepsilon_\nu \sim T^4 \sim a^{-4} \sim V^{-4/3}$$

not only for $q = 0$, but for any initial value of q. The parameter $|q|/T^3$ characterizes the degree of degeneracy of the neutrino-anti-neutrino gas (here T is the absolute temperature).

*Charge symmetry together with the condition of homogeneity (at least in the initial state) would lead to mutual annihilation of baryons and antibaryons; there are no forces which could separate the baryons and bring them together into such large aggregations as stars.

asymmetrical solution of symmetric equations with asymmetric initial conditions.

15. THE MASS OF THE CLOSED WORLD

It was emphasized above that at present it is not known whether or not the density ρ actually exceeds the critical value

$$\varrho_c = \frac{3H^2}{8\pi\varkappa}.$$

Therefore it is also unknown whether the world is closed. Without prejudging the answer to this question, it is interesting to consider one feature of the closed world: textbooks say that its mass is equal to zero, that the total energy and all the components of the momentum of a closed world are zero.

How is this assertion to be understood?

Here we must recall that there exists the concept of a gravitational mass defect.

In nuclear physics the mass of the deuteron is smaller than the sum of the masses of neutron and proton; the difference (defect) is

$$M_p + M_n - M_D = \frac{Q}{c^2},$$

where Q is the energy released when a proton and neutron combine into a deuteron.

In the same way the mass of a double star is smaller than the sum of the masses of the two individual stars (moving with the same speed) by the quantity $\kappa m_1 m_2 / r_{12} c^2$. The decrease in mass is equal to the (negative) gravitational interaction energy divided by c^2.

This decrease in mass can in principle be detected by measuring the gravitational field of a double star at a large distance.

In the case of a closed world the gravitational defect corresponding to the interaction of all of the stars and particles which compose it is exactly equal to the sum of the masses of all the stars, particles, and so on, taken separately, so that, roughly speaking,

$$M = \sum m_i - \frac{\varkappa}{2c^2} \sum_i \sum_k \frac{m_i m_k}{r_{ik}} = 0. \qquad (15.1)$$

This expression is a rough one, because the expression $-\kappa/r_{12}$ for the interaction energy of two unit masses is valid only for $r_{12} \ll a$, the radius of the world; the expression (15.1) conveys only the sense of the assertion; in the interaction with approximately homogeneous matter the main contributions are from distances $\sim a$, for which the interaction energy cannot be written simply.

The concept of the mass of a closed world is to some degree a mystical one, since there is no space external to this world, and no outside observer who could determine a gravitational field produced by the closed-in world.

With a different approach, however, the assertion that the mass of the closed world is zero acquires a quite definite meaning.

Let us consider a spherically symmetrical distribution of matter with a given density, i.e., with a given amount of matter in the invariantly defined unit volume inside some sphere. Farther out, beyond the surface of this sphere* there is empty space, and therefore at a sufficiently large distance the field is weak and we can uniquely determine the total mass of the distribution.

The amount of matter in our distribution is also quite uniquely defined; for example, we can speak of the number N of baryons (or the number N of stars). With this approach it turns out that the mass is proportional to N only for small N; the formula

$$M = mN$$

(m is the mass of one baryon or one star) is the first term of an expansion.

For small N, including the gravitational interaction, we have

$$M = Nm - \varkappa \frac{(Nm)^2}{R} = Nm - \text{const} \cdot (Nm)^{5/3}. \qquad (15.2)$$

If we do not confine ourselves to small values of N, then as N increases the mass M first increases, then goes through a maximum and after that begins to decrease! There are distributions of matter for which the addition of a new layer of matter, the addition of new particles to those already present, decreases the total mass of the system.

As N approaches a definite finite limit the mass M approaches zero, and this limit corresponds exactly to the case of the closed world.

16. STATEMENT OF THE PROBLEM OF THE INITIAL STATE

The theory of the expanding universe presents problems not only for observational astronomers and experimenters, but also for theorists.

The main problem is: it follows from the Fridman theory that there was a time when the density of matter in the universe was extremely large.

Will this conclusion remain valid when one takes into account the irregularities of the density distribution (galaxies, clusters of galaxies) and the existence of random velocities imposed on the Hubble distribution? The case of arbitrary initial conditions is considered in the following article by E. M. Lifshitz and I. M. Khalatnikov. Their investigation leads to the negative answer: in the most general case there cannot be any singularities (in particular, infinite density) in the solution.

This does not exclude the possibility of singulari-

*The concepts "farther" and "nearer" retain their full meanings in noneuclidean space when there is spherical symmetry.

ties in the initial condition* in the remote past of the universe.

If there has been a singularity, what was the physical state of the matter at that instant?

For the early stage, the Fridman theory gives as the law of change of density (see Sections 3 and 4)

$$\varrho = \frac{A}{t^2}, \tag{16.1}$$

where ρ is the density in g/cm^3, the time is in seconds, and on the simplest assumption the constant A is

$$A = 800\,000 \text{ g-sec}^2 \text{ cm}^{-3} = \frac{1}{6\pi\varkappa} \tag{16.2}$$

(κ is the Newtonian gravitation constant).

This law does not depend on whether the universe is closed or infinite; if the temperature is high, the constant A is changed slightly, A = 450,000 g-sec^2 cm^{-3}, and the form of the expression is unchanged. As can be seen from the formula, the time when $\rho = \infty$ is taken as $t = 0$.

Fifteen minutes after this instant, the density of the matter was equal to the normal density of water. Quite a number of questions naturally arise:

1) What did the matter consist of when its density (for $t \leq 10^{-5}$ sec) was larger than that of nuclear matter?

2) What state was the matter in, and what were its temperature and pressure?

3) What was there before the instant $t = 0$, for $t < 0$?

4) Can it be supposed that in the high-density state the matter was strictly homogeneous in space?

5) How did there arise from this state the present state of the universe, with a clearly marked nonuniform distribution of matter, concentrated in galaxies and clusters of galaxies?

These questions are of very different natures. For the third question there is at present not only no concrete answer, but no scientific approach to an answer. Perhaps some fusion of general relativity and quantum theory will make it possible to approach this question. Another possible point of view, however, is that the question itself is illegitimate and nonexistent, just as in the theory of relativity the question "which occurs earlier" does not exist for spatially separated events.

In this connection it must be emphasized that Fridman's work demonstrates the existence of a definite class of solutions of the equations of motion of the general theory of relativity with definite initial conditions; it has been shown that in its general features this solution describes the observed pattern of the world.

The form of the equations of motion of the general theory of relativity is to a great extent determined by general principles of theoretical physics. So far, however, there are no arguments which give a similar definite choice of the initial conditions (initial distributions of density and velocity) which lead to the Fridman solution.

To the simple question: "how do the enormous velocities of recession of distant galaxies arise?" the theory gives an evasive answer: there exists an initial velocity distribution at the instant of infinite density which leads at present to the observed velocities.

The first and second questions are quite concrete; they only need more precise formulation, to take account of the fact that particles can undergo transmutations, and at large densities the particles interact strongly and the concept of an individual particle loses its meaning. Therefore we must speak of the conserved quantities—the electric charge, the baryon charge or number, and the lepton charge (the importance of the last quantity has been pointed out independently by Saakyan).

These are the quantities that are strictly conserved in all laboratory experiments, in nuclear reactions, and in particular at the maximum energies in accelerators. It is natural to extend these laws and regard them as absolute, so as to apply them even up to $\rho = \infty$.

Moreover, the state of the matter is to be characterized not by the temperature, which changes in the course of the expansion, but by the entropy—more exactly, the specific entropy per baryon; this quantity is constant in an adiabatic expansion.

17. TYPES OF INITIAL CONDITIONS

We shall make more concrete our assumptions about the initial composition and state of the matter.

The first type of assumption that comes to mind is that of cold neutrons at ultrahigh density.

Since the time of Landau's classic work it is well known that such a state corresponds to a minimum energy at high density; the decay of even a small fraction of the neutrons, $n = p + e^- + \tilde{\nu}$ gives a state in which the electrons fill up all available levels up to a certain energy E_f, and further decay of neutrons is impossible. The energetically accessible states for an electron which might be produced in a decay are already occupied, and according to the Pauli principle a second electron cannot go into an occupied state.

What is obtained from the neutrons in the course of the expansion? As the density decreases, E_f also falls, and there is further decay of the neutrons. Since the lifetime of the neutron is about 10 minutes (when all electronic states are empty), the decay of the neutrons will occur at a density of the order of 1 g-cm^{-3}.

Just as it is produced, every proton will be subjected to extremely intense bombardment by the neutrons that have not yet decayed. The reactions

$$n+p=d+\gamma, \quad d+n=t+\gamma$$

*Moreover, it is not clear whether the additional condition of closedness for $\rho > \rho_c$ may also lead to singularities in the future (for $\rho < \rho_c$ the expansion will never give way to a contraction).

(d is the deuteron, t the triton) and the thermonuclear reactions

$$d+d,\quad d+t$$

lead to the disappearance of practically all of the protons, which are converted into deuterium and tritium.

Actually we know that in the initial stage of the evolution the matter was more than 90 percent hydrogen. Consequently, the assumption of cold neutrons is untenable.

A second, "hot" type of condition was suggested in 1950 by Gamow and his coworkers. It is assumed that the entropy is so large that with a radiation (photon) density of about 1 g/cm^3, i.e., with a huge temperature of the order of 7×10^8 degrees, the density of baryons was in the range 10^{-4}—10^{-8} g/cm^3. The point of this assumption is that when the baryon concentration is small there is less probability of collisions and nuclear reactions between them.

Originally Gamow wanted to use this sort of theory to obtain the abundances of the various chemical elements in nature.

In 1950 it was supposed that the Hubble constant is large (up to 500 km/sec per megaparsec), and would correspond to a time from the instant of the singularity to the present of less than 2×10^9 years, i.e., less than the age of the earth, for example. It could be believed that (apart from hydrogen and helium which have escaped) the composition of the earth is about that which arose during the expansion of the matter to the prestar stage.

Gamow's hypothesis could not be justified, since in the low-density hot matter there is no way to get across the nonexisting nuclei with atomic weights A = 5 and A = 8.

Now, with the long time scale $T \sim 10 \times 10^9$ years, the idea that nuclear synthesis has gone on in the stars and the elements formed have been thrown out in explosions of supernovae is a quite natural one. The earth was formed from matter which had undergone processing in stars at least once.

The prestar evolution of matter must have led to hydrogen as the initial substance for the construction of first-generation stars.

In Gamow's theory, with the low entropy (baryon density $\rho_1 = 10^{-4}$ g-cm^{-3} when the total density was $\rho = 1$ g-cm^{-3}) one gets something like 30 percent helium, which exceeds the helium content in old stars.

With a large entropy ($\rho_1 = 10^{-8}$ g-cm^{-3} with $\rho = 1$ g-cm^{-3}) one gets more than 90 percent hydrogen and less than 10 percent helium, which is a satisfactory composition.* Then, however, the theory leads to a large density of electromagnetic radiation at the present time. The present density of baryons is of the order of 3×10^{-31}; it is proportional to a^{-3}, and consequently since the time when ρ_1 was 10^{-8} and T was 7×10^8 degrees the baryon density has decreased by a factor 3×10^{22} so that a has increased by a factor 3×10^7 and T has decreased by the factor 3×10^7, and today we should have $T \sim 20$°K.

Along with this the energy density of radiation is 10^{-9} erg-cm^{-3} (1000 eV cm^{-3}), so that $\rho = \epsilon/c^2 = 10^{-30}$ g-cm^{-3}—larger than the density of the matter!

The heat capacity of the baryons is small, so that the atoms and molecules could not absorb any appreciable fraction of the radiation.

The idea of such a density of radiation is obviously in contradiction with both radio-astronomical observations and the indirect data of cosmic-ray theory. Therefore Gamow's theory must be rejected.

A type of theory is now being developed in which the prestar matter was cold (entropy = 0 at $\rho = \infty$) and consisted of a mixture of protons, electrons, and neutrinos (ν).

In the cold state the particles must be degenerate —they must occupy all of the states in momentum space up to the so-called Fermi energy E_f (which is different for different particles and depends on the density).

The E_f for neutrinos ($E_{f\nu}$) is larger than the corresponding quantity E_{fe} for electrons.

Therefore the presence of ν's suppresses the reaction

$$e^- + p = n + \nu$$

(the state the neutrino would have to go into is occupied).

The presence of ν's stabilizes the protons, as it were, at a high density. At a low density the protons are stable in themselves (when $E_{fe} < 1$ MeV, i.e., for $\rho < 10^8$ g-cm^{-3}).

Thus the new hypothesis leads to the conclusion that as the result of the initial, prestar period of expansion the ultradense matter would be converted into practically pure (99.99 percent) cold hydrogen. According to this conception all nuclear reactions occur only at a later state, when the stars are formed.

18. NONUNIFORMITY OF DENSITY AND ACCIDENTAL MOTION

At present discussions of the question of the limits of applicability of the Fridman solution are concentrated around three deviations from this solution, which are observed by astronomers.

These deviations are as follows:

a) The density distribution is nonuniform: the greater part of the matter is concentrated in the stars, and these in turn are not distributed randomly through space, but are concentrated in galaxies, and in turn these form clusters of galaxies; perhaps we should end this clause not with a period, but with the words "and so on"—beyond the clusters there are super-

*It is true that one still gets a few percent of deuterium, which is in contradiction with the observations.

clusters, and there is no assurance that there are no still huger structural units;

b) There are accidental velocities; the idea of all objects moving with velocities $\mathbf{u} = H\mathbf{r}$ is a decided idealization. Actually the velocities differ considerably (by 200—500 km/sec) from this law. If we write

$$\mathbf{u} = H\mathbf{r} + \mathbf{w}, \tag{18.1}$$

we can call $\mathbf{w}$ the accidental velocity.

What is actually measured is only the longitudinal component of the velocity (by the Doppler effect); the distances of many distant objects are measured with extremely small accuracy. Therefore we can make judgments about the accidental velocities of distant objects only if there is reason to think that several objects belong together (and do not just accidentally appear in the same direction), while their velocities are different. Thus the evidence about accidental velocities is very scanty.

The question of accidental velocities is a sharp one, because in the homogeneous Fridman solution there is a theorem on the damping out of any accidental velocity in the expanding universe. This theorem is in essence the same as the conclusion about the red shift of frequencies; in the course of time the momentum of a particle falls off in inverse proportion to the radius of the universe,

$$p = p_0 \frac{a_0}{a},$$

$$\frac{dp}{dt} = -\frac{p}{a}\frac{da}{dt} = -pH, \tag{18.2}$$

where H is the Hubble constant.

Furthermore it can be said that the red shift is a special case of the damping of accidental motions as applied to a particle—the light quantum: the momentum of the photon is $\hbar\omega/c$, proportional to its frequency. Hoyle points out that at present there are accidental motions of galaxies with speeds of the order of 200 km/sec, and poses the question: what was the accidental velocity at the time when the galaxies separated from each other? Since the size of galaxies is about a fiftieth of the average distance between them, the radius of the universe has increased by a factor 50 since they separated.

According to Hoyle this means that at the time of separation the velocities were $\sim 10{,}000$ km/sec, which is absurd, as it corresponds to too large a kinetic energy.

The mistake in this argument is that the damping of the velocity is regarded as if one particular selected galaxy were moving through a strictly homogeneous distribution of matter. Actually the two factors (the nonuniformity of the density and the accidental velocities) are closely connected.

In the gravitational field of nonuniformly distributed matter the law of decrease of accidental velocities does not hold at all.

We can give a most extreme, popular example: the revolution of the earth around the sun is a special case of an "accidental" motion. It is obvious that the general Hubble expansion has no effect at all on this motion; the earth's orbit does not expand, and its speed of motion in its orbit does not decrease in the course of time.

This is due to the fact that on the scale of the earth's orbit the existence of the sun is a huge violation of the uniformity of the distribution of matter; the ratio of the mass of the sun to $4\pi/3R^3$, where R is the radius of the orbit, is of the order of 10^{-7} g/cm^3, larger than the mean density in the universe by a factor 10^{23}.

In the general case N. A. Dmitriev and the writer have succeeded in deriving a curious estimate: if in the initial state the inhomogeneity of the density and the random velocities were small, then owing to the gravitational interaction the kinetic energy of the accidental motion can come to be of the order of (but not greater than!) the quantity

$$\frac{\varkappa}{2}\int\int \frac{(\varrho(\mathbf{r}_1)-\bar{\varrho})(\varrho(\mathbf{r}_2)-\bar{\varrho})}{r_{12}}\,dV_1\,dV_2. \tag{18.3}$$

In matter which is uniform on the average, the distant regions (large values of r_{12}) make vanishingly small contributions to the integral, and therefore we can use the Newtonian approximation and the Newtonian potential.

There is no space here for the exact formulation of the theorem and the conditions under which it holds. Its meaning is that the appearance of inhomogeneities in the density is accompanied by conversion of gravitational energy into kinetic energy of accidental motions.

In the real universe with its nonuniform and nonconstant density there is no theorem of the smooth decrease of all velocities with time, or of the damping out of accidental motions. At present there are no data which could indicate definitely whether the speeds of accidental motions exceed the values given by the estimate we have stated.

When stars are being formed, moreover, energy from nuclear reactions can be released in addition to gravitational energy. There have so far been no studies at all of the possibility that this energy may be partially converted into the energy of accidental motions, but it must not be forgotten.

19. GRAVITATIONAL INSTABILITY

When uniformly distributed matter gets bunched together into separate aggregations gravitational energy is released and converted into kinetic energy. The resulting motion in turn intensifies the nonuniformity of the density. Uniformly distributed matter is in an unstable state in relation to the action of gravitational forces.

At the beginning of the century this instability was studied mathematically by Jeans. He considered a gas with a definite pressure and speed of sound. Under small-scale perturbations with short wavelengths the pressure gradient is large, so that the pressure differences smooth out the nonuniform density. For large-scale perturbations the pressure is of no importance. The time variation of every perturbation follows the law

$$F = Ae^{\omega t} + Be^{-\omega t}, \tag{19.1}$$

where A and B can be expressed in terms of the initial quantities. For large-scale processes the characteristic quantity ω is given by

$$\omega = \sqrt{4\pi\varkappa\varrho}. \tag{19.2}$$

Jeans considered perturbations applied to stationary matter which is uniform in space and constant in time. We saw in Sections 3 and 4 that matter at rest and constant in time does not satisfy even the Newtonian equations. The article by Lifshitz and Khalatnikov gives a detailed description of the exact solution of the problem of the development of small perturbations imposed on the Fridman solution which describes an expanding homogeneous universe.

The general ideas about the physics of the process are essentially quite unchanged: the matter is more strongly attracted toward the region where the perturbation has caused an increased density—this is the cause of the instability.

In the expanding universe $\bar{\rho}$ is not constant, but changes in the course of time. Therefore ω also changes. In the exponent we naturally replace the product ωt with the integral $\int \omega\, dt$: at each instant the perturbation is increasing in accordance with the instantaneous value of the characteristic quantity ω, which depends on the density.

In the initial period,

$$\varrho = \frac{1}{6\pi\varkappa t^2}, \qquad a(t) = \text{const} \cdot t^{2/3}. \tag{19.3}$$

Substituting, we get

$$\omega = \frac{1}{t}\sqrt{\frac{2}{3}}, \quad \int \omega\, dt = \sqrt{\frac{2}{3}} \ln t + C,$$

$$e^{\int \omega\, dt} = \text{const} \cdot t^{\sqrt{\frac{2}{3}}} = \text{const} \cdot a^{\sqrt{\frac{3}{2}}}. \tag{19.4}$$

Lifshitz' exact solution is

$$F = Aa + Ba^{-3/2}, \tag{19.5}$$

and the exponents $+1$, $-\frac{3}{2}$ do not differ much from $\pm(\frac{3}{2})^{1/2}$.

That there are two terms, one increasing and one decreasing, in the Jeans solution and in the Lifshitz solution, is typical of unstable mechanical systems.

Let us consider the motion without friction of a heavy material point P near a smooth maximum O, which is obviously a position of unstable equilibrium. Obviously we here also get for the coordinate of the point

$$x = Ae^{\omega t} + Be^{-\omega t}.$$

In principle we can prescribe the initial position and velocity (directed toward the maximum height) so that the point will rise, moving more and more slowly, and stop just at the top; this is indeed the particular solution

$$A = 0, \quad x = Be^{-\omega t}.$$

With arbitrarily assigned position and velocity, however, obviously $A \neq 0$, and in the course of time, in the future, the term $Ae^{\omega t}$ will always come to predominate—the perturbation will grow. This also applies to the case of the distribution of matter in the universe.

What can be said about the past?

In our analogy (Fig. 5), if we see the point P somewhere to one side of O and do not know its velocity, we cannot say whether it got to its position by being thrown up from somewhere below, or has rolled down from the position of equilibrium.

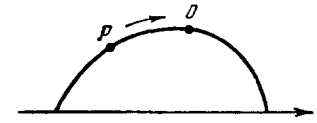

FIG. 5

In just the same way, it is possible to suppose that the nonuniformity in the universe observed at present is the result of some larger nonuniformity in the past, which is decreasing according to the law

$$F = Ba^{-3/2}.$$

This assumption about the past, however, although possible in principle, is improbable and implausible. If in the past there were perturbations of both types, increasing (Aa) and decreasing ($Ba^{-3/2}$), then it is the increasing ones that remain now, and the present nonuniformity is larger than it was in the past.

There is no possibility of our directly observing a change with time of the distribution of the galaxies. Therefore only careful measurements of the velocity distribution can in the future give an objective answer as to the nature of the observed nonuniformity (increasing or decreasing).

During the time in which the density of matter has changed from nuclear density (10^{14} g-cm^{-3}) to the present value (3×10^{-31} g-cm^{-3}) the radius a has increased by the factor $(10^{14}/3 \times 10^{-31})^{1/3} = 10^{15}$. From the time when $\rho = 0.1$ g-cm^{-3} to the present, a has increased by the factor 10^{11}. Therefore even a trifling inhomogeneity of the order $\delta\rho/\rho \sim 10^{-9}$ at the time when the density was of the order of 0.1 g-cm^{-3} would be enough so that the gravitational instability would

have led to a strong inhomogeneity (of the order of unity) in the distribution of matter as observed now.

Owing to the gravitational instability it is quite permissible to suppose that the Fridman equations held exactly, that the matter was exactly uniform in the early stages of the evolution.

The question of the (small) initial amplitudes of inhomogeneities which are needed for the development of the instability has at present only been stated.

The density fluctuations in an ideal gas consisting of separate independent molecules, when considered on the scale of the stars, are completely inadequate.

Some part may have been played by phase transitions, which must occur in the course of the expansion of cold hydrogen (conversion of metallic into molecular hydrogen, formation of the gas phase from liquid hydrogen). It may be that extremely small inhomogeneities existed initially, but then the theory becomes arbitrary.

20. CONCLUSION

The test of time is the strongest and unerring test of a scientific theory.

The cosmological theory of the expanding universe put forward by A. A. Fridman has already been undergoing this test for 40 years; in the twentieth century, when the development of science is enormously accelerated, 40 years is the equivalent of several centuries in the past.

From this test the Fridman theory has emerged strengthened. Observations have confirmed the actual fact that the universe is not stationary. Repeated attempts to find some other explanation of the red shift of spectral lines have failed ingloriously.

Theories that try to combine the recession of the nebulas with the preconceived idea of a stationary universe by giving up all of the laws of physics* are in their death throes.

The difficulties of reconciling the short time scale with data on the age of the earth have vanished with the more accurate values of distances, which have led to a decrease in the value given to the Hubble constant.

There are many unsolved problems in cosmology, but their solutions are to be sought on the basis of Fridman's theory, in the framework of the general ideas he developed.

A survey report by Wheeler at the Solvay Congress in 1958 gives a good description of the scientific drama of cosmology.

"Past history warns us of the danger of neglecting Einstein's theory when it comes into collision with preconceived ideas. He himself (Einstein) tells us how unhappy he felt when the general theory of relativity predicted that a world of finite density must be changing in size; how he invented a new artificial term containing a "cosmological constant" in order to cancel out this "unreasonable" change of size; about the subsequent discovery that the world is actually expanding; and about his conclusion that the cosmological term should never have been introduced at all; that conclusions drawn from a simple, directly and consistently developed theory must be taken seriously."

Only one thing is not stated here—that the correct solution came from the Soviet Union and was due to A. A. Fridman.

A favorite saying of A. A. Fridman was: "The waters into which I am stepping have not yet been crossed by any man."* In his brief notes concerning his cosmological solutions Fridman not only himself stepped into a new domain, but also showed to us fruitful ways onward and forward, into the unknown.

Notes added in proof. To page 482: Nonuniformity of the density distribution in the universe decidedly alters these predictions. In particular, if by chance there is no matter inside the cone of rays joining the outline of the object to the point where the observer is, the angular diameter θ of the object is smaller. In this case θ does not have a minimum [see an article by the writer, Astron. zhurn. (1963, in press)].

To page 484: The latest data evidently indicate that the temperature of intergalactic thermal radiation is below 1°–0.5°K, so that the energy density is smaller than stated in the text by a factor 10^5. Further improvements in accuracy and lowering of the value for the temperature will require measurements with apparatus located outside the earth's atmosphere.

*A detailed criticism of the theory of the spontaneous creation of matter, which aspires to replace the Fridman theory, is given in the author's article mentioned at the end of Section 1.

*The words of Dante Alighieri: "L'acqua ch'io prendo giammi non si corse"; I quote from the reminiscences of E. P. Fridman [Geofizicheskiĭ sbornik 5, No. 1 (1927)].

Translated by W. H. Furry
28