THE INITIAL STAGE OF AN EXPANDING UNIVERSE AND THE APPEARANCE OF A NONUNIFORM DISTRIBUTION OF MATTER

A. D. SAKHAROV

Submitted to JETP editor March 2, 1965

J. Exptl. Theoret. Phys. (U.S.S.R.) 49, 345-358 (July, 1965)

A hypothesis of the creation of astronomical bodies as a result of gravitational instability of the expanding universe is investigated. It is assumed that the initial inhomogeneities arise as a result of quantum fluctuations of cold baryon-lepton matter at densities of the order of 10^{98} baryons/cm^3. It is suggested that at such densities gravitational effects are of decisive importance in the equation of state and the dependence of the energy density ϵ on the baryon density n can qualitatively be described by graphs a of b of Fig. 1. ϵ vanishes at a certain density $n = n_0$. A theoretical estimate (containing some vague points) yields initial inhomogeneities in the distribution of matter which can explain the origin of clusters of 10^{62}—10^{63} baryons (10^5—10^6 $M_\odot$). The calculated mass is smaller than that of the galaxies by a factor of 10^5—10^6; it is in fact closer to the masses of globular clusters. The hypothesis is proposed that galaxies are produced as a result of an increase of nonuniformities in the motion and distribution of the gas formed during gravitational collapses of the primordial stellar clusters. According to this hypothesis the plane component of the galaxy is produced from the gas, whereas the spherical component consists mainly of clusters of primordial stars captured by the gravitational field of the rotating gas cloud.

UNITS AND NOTATION

In the gravitational system of units we put $\hbar = c = G = 1$.

The unit of length is $c^{-1/2}\hbar^{1/2}G^{1/2} = 1.61 \times 10^{-33}$ cm.

The unit of time = $c^{-3/2}\hbar^{1/2}G^{1/2} = 5.35 \times 10^{-44}$ sec $= 1.7 \times 10^{-51}$ year.

The unit of mass = $c^{1/2}\hbar^{1/2}G^{-1/2} = 2.18 \times 10^{-5}$ g $= 10^{-38}$ $M_\odot$.

The baryon density is n; $n = 1$ at a density 0.24×10^{99} baryon/cm^3.

The average distance between baryons is $a = n^{-1/3}$. The relative perturbation of the density is expanded in a Fourier series of the orthogonal functions Φ_κ, defined (to simplify the normalization) in a cube of the volume period $V = V_0 a^3$ (V_0 does not depend on the time):

$$\frac{\Delta n}{n} = \sum_{\kappa} \Phi_\kappa(\xi) z_\kappa(t). \qquad (1)$$

Here $\xi = x/a$ is the co-moving dimensionless spatial coordinate (3-vector), and κ is a dimensionless wave vector (independent of the time). The function Φ satisfies the normalization condition and the equation

$$\int_{V_0} |\Phi|^2 d\xi = 1, \qquad \Delta\Phi + \frac{\kappa^2}{a^2}\Phi = 0.$$

For example, $\Phi = v_0^{-1/2} \exp(i\kappa\xi)$. The quantity z_κ defined in accordance with (1) is called the "amplitude of the inhomogeneity."

ϵ—energy density in the rest frame of the material.

$p = n d\epsilon/dn - \epsilon$ is the pressure.

A dot over a letter denotes a time derivative.

C = the limit of $(\dot{a})^2$ as $t \to \infty$; the symbol $\sim$ stands for equal in order of magnitude; the symbol $\propto$ denotes proportionality.

1. INTRODUCTION

The cosmological theory of the expanding universe is at present generally accepted. This theory is based on a nonstationary solution obtained by A. A. Friedmann for Einstein's equations of general relativity, and explains, in particular, phenomenon of the "red shift" (see, for example, the series of articles devoted to the memory of A. A. Friedmann [1]).

Recently Ya. B. Zel'dovich (see, for example,[1]) presented convincing arguments in favor of the assumption that matter was cold in the initial, dense state and indicated that under certain assumptions concerning the initial relations between the baryon and lepton densities it is possible to explain, within the framework of these ideas, the predominant abundance of hydrogen in the uni-

verse and the low temperature of intergalactic space. It can be assumed that during the early stage of the expansion the matter in the universe was practically homogeneous, and the "primordial' astronomical objects were the result of gravitational instability. Although many astronomers and astrophysicists object to this point of view, an investigation of this point of view is essential. For the development of such a hypothesis, great significance is attached to a study of the laws whereby small inhomogeneities of density build up, and to the determination of the statistical characteristics of the initial inhomogeneities. The first problem was solved within the framework of the theory of the expanding universe by E. M. Lifshitz and again considered by Ya. B. Zel'dovich (see [1,2]), while the solution of the second problem calls for an examination of the initial stage of expansion of the universe, and therefore contains many hypothetical premises. However, by combining theoretical considerations and astronomical observations we can hope to obtain the necessary initial data for the solution of cosmogonical problems and information concerning the physical laws obtaining at extremely high densities of matter.

The fundamental equation of the theory of the expanding universe can be taken to be one of Einstein's equations

$$R_0^0 = 8\pi(T_0^0 - {}^1/_2 T).$$

Substituting for the case of a homogeneous world $R_0^0 = -3\ddot{a}/a$ and the components of the energy-momentum tensor in the form $T_0^0 = \epsilon$, $T = \epsilon - 3p$, we get

$$\ddot{a} = -\frac{4\pi}{3}(\varepsilon + 3p)a. \tag{2}$$

Equation (2) has the integral

$$\dot{a}^2 = \frac{8\pi}{3}\varepsilon a^2 + C, \tag{3}$$

where the integration constant C determines, as is well known, the sign and the magnitude of the spatial curvature. If we take Oort's estimate for the average density of matter at the present time $\bar{\rho} = 3 \times 10^{-31}$ g/cm^3, then $\dot{a}^2 > 8\pi\epsilon a^2/3$ and the constant $C > 0$, although it is very small: $C \approx \dot{a}^2 = 1.6 \times 10^{-52}$. This is a case of Lobachevskiĭ geometry. However, estimates of $\bar{\rho}$ are not very accurate and it is not excluded that $C = 0$ and $\bar{\rho} = 1.2 \times 10^{-29}$ g/cm^3, and that at the present time $\dot{a}^2 = 2 \times 10^{-53}$ ($\dot{a}^2 \to 0$ as $t \to \infty$), and even $C < 0$ is not excluded.

It is assumed in this paper that the equation of state of matter at $n \ll 1$ coincides qualitatively with that for a degenerate Fermi gas, and that the effect of interaction and transformation of fermions does not play a very important role. Accordingly, $\epsilon \sim n^{4/3}$ for $n^{1/3} > M$ ($M = 0.76 \times 10^{-19}$ is the baryon mass) and $\epsilon = Mn$ when $n^{1/3} < M$.

We distinguish arbitrarily between four (or three) stages of the expansion of the universe:

stage 1	$\varepsilon \lesssim 1$	$a \lesssim 1$	$1 \gtrsim t$
stage 2	$\varepsilon \sim n^{4/3}$	$a \sim t^{1/2}$	$1 < t < M^{-2}$
stage 3	$\varepsilon = Mn$	$a \sim t^{2/3}M^{1/3}$	$M^{-2} < t < MC^{-3/2}$
stage 4	$p \ll \varepsilon$	$a = C^{1/2}t$	$MC^{-3/2} < t.$

During the first stage the gravitational interaction of neighboring particles is ~ 1 and has an important influence on the equation of state. An investigation of this stage is essential for the determination of the initial inhomogeneities of quantum character. If $C = 0$ then the fourth stage does not set in.

Investigations carried out by E. M. Lifshitz [2] and Ya. B. Zel'dovich (verbal communication) of the laws governing the build up of small density perturbations can be summarized as follows. Neglecting the effects of the spatial pressure gradient, the perturbations build up during all the stages in inverse proportion to $\dot{a}^2$:

$$z_\varkappa(t) = z_{0\varkappa}/\dot{a}^2, \tag{4}$$

where $z_{0\kappa}$ is the "initial" amplitude at $\dot{a} \sim 1$.

Ya. B. Zel'dovich called our attention to the importance of taking into account the pressure effects, $p \sim \epsilon$, during the second stage for short waves. According to Zel'dovich, the stability condition is $M^{-2} > t > \kappa^{-2}$ (that is, in particular, $\kappa > M$). If the stability condition is satisfied, oscillations of the particle density take place with conservation of the adiabatic invariant.

When $\kappa \gtrsim M$, formula (4) for the third and fourth stages should be replaced by a formula that takes into account the effect of the pressure gradient in the second stage:

$$z_\varkappa(t) = z_{0\varkappa}B(\varkappa)/\dot{a}^2, \quad t > M^{-2}. \tag{5}$$

Calculation of the function $B(\kappa)$ for $\kappa \sim M$ calls for the knowledge of the equation of state when $n \sim M^3 = 10^{42}$ baryon/cm^3.

We accept in this paper Zel'dovich's hypothesis that the initial temperature of matter in the universe was zero, and concentrate our attention on a consideration of quantum fluctuations of the density.

We note that during the course of the expansion of the universe the temperature of the matter first increases somewhat, owing to the incomplete adiabaticity of the transformation of the elementary particles in the second stage, after which it drops

as a result of the adiabatic expansion of matter, reaching $\sim 10^{-7}$ erg or less at densities on the order of 10^{42} baryon/cm^3. Zel'dovich's estimates (verbal communication) show that the temperature effect is not significant in the formation of inhomogeneities.

2. FUNDAMENTAL EQUATIONS FOR SMALL DENSITY PERTURBATIONS

The equations derived in this section coincide in content with the equations of E. M. Lifshitz,[2] but are written in a different coordinate system, with a view of going over to quantum theory. The quantization of small oscillations is easiest to carry out if the equations of motion are written in Lagrangian form with

$$L_\varkappa = {}^1/_2 m(t)\dot{z}^2 - {}^1/_2 r(t) z^2.$$

We then introduce a wave function $\Psi_K(z, t)$, satisfying the Schrödinger equation [m(t) and r(t) are known functions]

$$-\frac{1}{2m(t)}\frac{\partial^2\Psi}{\partial r^2}+\frac{r(t)z^2}{2}\Psi = -i\frac{\partial\Psi}{\partial t}. \tag{6}$$

The "mass" m should correspond here to the value (7), obtained for short-wave oscillations (that is, for an unperturbed metric) as a result of a simple direct calculation of the change in the fraction of the Lagrange function

$$L_m = \int \sqrt{-g}\,\varepsilon\, d\xi,$$

connected with the matter. Expanding in powers of z^2 and $\dot{z}^2$, we get

$$m_\varkappa(t) = \frac{a^2}{\varkappa^2}\frac{d\varepsilon}{dn}, \quad r_\varkappa(t) = n\frac{d^2\varepsilon}{dn^2} = \frac{dp}{dn}. \tag{7}$$

To obtain r in the case of long-wave perturbations, it is best to start from the equations of motion, obtaining the latter by a method which is close to the method of Ya. B. Zel'dovich (verbal communication).

Let us consider the expansion (1) of an arbitrary perturbation in orthogonal functions Φ, satisfying the equation

$$\Delta\Phi + (\varkappa/a)^2\Phi = 0. \tag{8}$$

The function Φ can be written in the form of an exponential or a sinusoid, or else in spherically symmetrical form

$$\Phi \sim \sin \varkappa\xi / \varkappa\xi$$

(ξ is the radius). The dependence of z_K on t is determined in all cases only by the value of the parameter κ in Eq. (8). To find this dependence, it is sufficient to satisfy the equations of motion at the point $\xi = 0$ in the spherically symmetrical case. Then, by virtue of the isotropy of space, the equations of motion will be satisfied automatically in all the remaining points. The fundamental Eq. (2) is the equation of motion at the point $\xi = 0$, accurate to effects of the spatial pressure gradient.

For long-wave perturbations we can neglect the effects of the spatial pressure gradient. We shall denote by the index zero the unperturbed quantities and by the index 1 infinitesimally small quantities of first order

$$a = a_0 + a_1, \quad t = t_0 + t_1, \quad \varepsilon = \varepsilon_0 + \frac{d\varepsilon}{da}a_1 \text{ etc.} \tag{9}$$

Substituting (9) in (2) and separating the terms of first order of smallness we obtain equations for the perturbation amplitude. Here we consider t as the local time [dt = ds at the point $\xi = 0$; it is just this t which enters in (2)], and t_0 is regarded as the "world time" of the coordinate system.

Lifshitz and Zel'dovich have assumed that $t_1 = 0$. Zel'dovich actually used variation of the integration constants in the analytic solution of (2), a simple matter if one uses the special equation of state $\epsilon \sim n^\gamma$; he considered $\gamma = 1$, $\gamma = 4/3$, and $\gamma = 2$. In the general case, the substitution (9) with the condition $t_1 = 0$ in (2) leads to the equation of motion

$$\frac{1}{a_0}\frac{d}{dt}\left(a_0^2\frac{d}{dt}\left(\frac{a_1}{a_0}\right)\right) = 4\pi\left(n\frac{d\varepsilon}{dn}+3n\frac{dp}{dn}\right)a_1. \tag{10}$$

We note that inasmuch as one of the solutions of this equation should be $a_1 \sim \dot{a}_0$ [a time shift of the unperturbed solution of (2)], Eq. (10) can be written in the form

$$\ddot{a}_1/a_1 = \dddot{a}_0/\dot{a}_0. \tag{11}$$

Introducing the relative amplitude of the inhomogeneities

$$z = n_1/n_0 = -3a_1/a_0, \tag{12}$$

we obtain from (10) or (11) for the special equation of state $\epsilon \sim n^\gamma$ Zel'dovich's result, namely, two independent solutions

$$z \sim t^\alpha, \quad \alpha_1 = 2 - 4/3\gamma, \quad \alpha_2 = -1.$$

We now forego the condition $t_1 = 0$, imposing an additional condition, which is useful to us, $m = (a/\kappa)^2\, d\epsilon/dn$ [as in (7)]. The first term in an equation of the type (10) should be of the form ($\omega \equiv d\epsilon/dn$)

$$\frac{1}{a_0}\frac{d}{dt_0}\left(\omega a_0^2\frac{d}{dt_0}\left(\frac{a_1}{a_0}\right)\right).$$

Obviously t_1 must be chosen such as to add to the equation a term

$$a_0 \frac{d\omega_0}{dt} \frac{d}{dt}\left(\frac{a_1}{a_0}\right).$$

We put

$$\dot{t}_1 = -\frac{1}{\omega}\frac{d\omega}{dt}\left(\frac{da_0}{dt}\right)^{-1} a_1 = -z\frac{dp}{d\varepsilon}. \qquad (13)$$

Recognizing that, accurate to second-order terms,

$$d^2a/dt^2 = \ddot{a}_0 + \ddot{a}_1 - 2\dot{t}_1\ddot{a}_0 - \ddot{t}_1\dot{a}_0,$$

we obtain under the gauge condition (13):

$$\frac{1}{a^2\omega}\frac{d}{dt}\left(\omega a_0^2\frac{dz}{dt}\right) = \left\{4\pi n\left(\frac{d\varepsilon}{dn} + 3\frac{dp}{dn}\right) - 8\pi\frac{dp}{d\varepsilon}(\varepsilon + 3p) + 3\frac{\dot{a}_0}{a_0}\frac{d}{dt}\left(\frac{dp}{d\varepsilon}\right)\right\} z. \qquad (14)$$

[Starting with (14) we shall omit the subscript zero in differentiating with respect to the world time t_0.]

For the special equation of state $\epsilon = n^\gamma$ we obtain again two linearly-independent solutions

$$z \sim t^\alpha, \quad \alpha_1 = 2 - 4/3\gamma, \quad \alpha_2 = 1 - 2/\gamma.$$

The second solution vanishes (for long-wave perturbations) when the reference frame is transformed to t_1 = const. To the contrary, when $\dot{t}_1 = -zdp/d\epsilon$, the solution with $\alpha_2 = -1$ vanishes.

For the short-wave oscillations, Eq. (14) should be supplemented by a term which takes into account the spatial pressure gradient [according to formula (7)]. Taking (3) into account, the equation takes (for the particular case C = 0) the form

$$\frac{1}{a^2\omega}\frac{d}{dt}\left(a^2\omega\frac{dz}{dt}\right) = \left\{4\pi n\left(\frac{d\varepsilon}{dn} + 3\frac{dp}{dn}\right) - 8\pi\frac{dp}{d\varepsilon}(\varepsilon + 3p) - 24\pi\varepsilon n\frac{d}{dn}\left(\frac{dp}{d\varepsilon}\right) - \frac{\varkappa^2}{a^2}\frac{dp}{d\varepsilon}\right\} z. \qquad (15)$$

When γ = const, the solution of this equation is expressed in terms of Bessel functions; for example, when $\gamma = 4/3$ we have increasing and decreasing solutions of the form ($\vartheta \sim t^{1/2}\varkappa$)

$$z \propto \begin{cases} \cos\vartheta - \vartheta^{-1}\sin\vartheta, \\ \sin\vartheta + \vartheta^{-1}\cos\vartheta. \end{cases}$$

3. THE FUNCTION B(κ) AND THE MASSES OF STARS OF PREGALACTIC ORIGIN

Yu. M. Shustov and V. A. Tarasov have at our request solved Eq. (15), with the aid of an electronic computer, for different values of κ. The calculations were made for the simplest equation of state, satisfying $\epsilon = nM$ with $n^{1/3} \ll M$ and $\epsilon = An^{4/3}$ with $n^{1/3} \gg M$ (A is a constant ~ 1)

$$\varepsilon = n(M^2 + A^2n^{2/3})^{1/2}. \qquad (16)$$

The asymptotic value of p as $n \to 0$ is less satisfactory in this equation; for $n^{1/3} > M$ no account was taken of the transformations and interactions of the baryons.

The function a(t) can be obtained in the case of Eq. (16) analytically (Shustov). Shustov and Tarasov find, by integrating (15) the limiting value as $t \to \infty$ of the auxilliary variable

$$\zeta = z(1 + a^2M^2/A^2)^{-1/2},$$

putting $d\zeta/dt = dz/dt \sim z_0$ as $t \to 0$. It is obvious that $\zeta(\infty) \backsim z_0B$.

In accordance with the results of the sections that follow, we put $z_0 \sim \kappa$. $\zeta(\infty)$ is a function of the parameter $A^{1/2}\kappa$. This function is oscillating and sign-alternating, but attenuates rapidly with increasing κ.

The distribution of the stars over the masses depends in a complicated nonlinear manner on z, and can hardly be obtained without very cumbersome calculations. For a qualitative estimate let us determine the function F(N) = ΔN/N, where N is the average number of particles in a certain volume V = N/n and ΔN is the mean-square deviation

$$(\Delta N)^2 = \overline{[N(V) - \bar{N}(V)]^2}.$$

We obtain F(N) from formulas which, according to N. A. Dmitriev (verbal communication from Zel'dovich) can be employed when F(N) is of such form that it is essential to exclude the effect of the fluctuations on the boundary of the averaging region. In our notation

$$F(N) = \left(\int_0^\infty z^2(\varkappa)\varkappa^2 e^{-\varkappa^2/4\alpha^2}d\varkappa \Big/ N\int_0^\infty \varkappa^2 e^{-\varkappa^2/4\alpha^2}d\varkappa\right)^{1/2},$$

$$N = 4\pi\int_0^\infty e^{-\alpha^2\xi^2}\xi^2 d\xi, \qquad \text{i.e.} \quad \alpha = \pi^{1/2}/N^{1/3}. \qquad (17)$$

The intuitive meaning of formulas (17) is that ΔN and N are calculated for a volume V with diffuse boundaries. For statistically independent particles we have z = 1 and $F(N) = N^{-1/2}$.

It is obvious that the formation of objects with N particles is possible only if $F(N) \sim 1$. It can be assumed that if there exist maxima of the plot of the distribution of the stars by masses, then they correspond at least approximately to the maxima of functions of the form $N^m F(N)$ ($m \sim 1$). Putting (arbitrarily) for estimating purposes $m = 1/2$, we obtain, using the results of Shustov and Tarasov (assuming A = 5.8, corresponding to a mixture of 1 proton + 1 electron + 1 neutrino

without account of the transformation of the baryons and of the interactions) a maximum of the function $N^{1/2}F$ at $N = 0.47 \times 10^{57}$, that is, at a mass $m = 0.4M_\odot$. The assumption that the stars most preferably produced are those with mass $\sim M_\odot$ is due to Zel'dovich, who started from an examination of the boundaries of the stability region.

One could have assumed that the stars which enter in globular clusters were produced at an earlier stage of the expansion of the universe, and their mass distribution should be described by the theory developed above. Sandage [3] investigated the distribution of the stars of the globular cluster M3 with respect to their absolute stellar magnitudes. Maxima were obtained near $0.8M_\odot$ and at $(3-4)M_\odot$. Whereas the first maximum could be related somehow with the foregoing theory, after refining the equation of state and the averaging laws, the presence of the second mass maximum (which consists essentially of the variable stars of type RR-Lyrae), cannot be explained. It is obvious that stars brighter than the sun cannot be of the same age as the universe, and at least part of their mass is of secondary origin.

The function $F(N)$ for arbitrary $z(\kappa)$ increases monotonically with decreasing N (remark by Shustov). It is possible that this means that the great part of the primary stars has a mass smaller than $0.4M_\odot$ (determined by phenomena which we did not take into account), and all the masses of Sandage's distribution are of secondary origin.

4. THE HYPOTHESIS CONCERNING THE EQUATION OF STATE WITH $n \sim 1$

In the case of a degenerate Fermi gas of relativistic noninteracting particles, $\epsilon = An^{4/3}$ with $A \sim 1$. The electromagnetic interaction of the extremely relativistic fermions in all orders of the perturbation-theory series is $\backsim n^{4/3}$, that is, it only changes the value of the coefficient A. We assume that an account of the mutual transformation of fermions and of all possible interactions still leaves us with a formula of the type

$$\varepsilon = A(n)n^{4/3}, \quad \text{where } A(n) \sim 1 \qquad \text{for } 1 > n^{1/3} > M.$$

Another assumption of the theory is extrapolation of the concepts of general relativity down to distance scales of the order of 1.61×10^{-33} cm. In all probability, at such scales our concepts of space should be reviewed, that is, we are on the border of applicability of the existing theory, a fact which does not exclude the possibility of obtaining qualitatively correct results in the sense of the "correspondence principle."

At densities on the order of unity (in gravitational units) the exchange and correlation gravitational interactions of fermions become comparable in order of magnitude with the Fermi energy. Therefore the gravitational interactions become decisive in the equation of state. Qualitatively the role of these effects is obvious from the expansion of ϵ in powers of the gravitational constant [more accurately, in powers of $(G\hbar/c^3)n^{2/3}$]:

$$\varepsilon = An^{4/3} + Bn^2 - Cn^{8/3}. \tag{18}$$

In this expansion, which is valid when $n \ll 1$, the second term is the exchange gravitational interaction, and the third is the correlation interaction. A, B, and C are of the order of unity and are larger than zero. When $n \sim 1$ the kinetic and exchange energies are as before greater than zero, and the correlation energy is as before less than zero, but owing to the effect of the subtraction of quantitatively unknown expressions, any theoretical estimates of even the qualitative nature of the curve $\epsilon(n)$ are hypothetical. We assume that a dependence analogous to (18) takes place also for $n \sim 1$, that is, that for a certain value $n = n_\alpha$ the quantity ϵ reaches a maximum, and that $\epsilon = 0$ when $n = n_0$ (see the Figure, curve a).

An equation of state of this form makes it possible to formulate correctly the problem of calculating the initial perturbations (see below).

More satisfactory is also the qualitative character of the unperturbed solution. The dependence of the world radius a on the time is obtained from (2). For $t = 0$ we put $\dot{a} = 0$, i.e.,

$$\varepsilon = -3C/8\pi a^2, \quad \ddot{a} \sim -(\varepsilon + 3p) \approx -3nd\varepsilon/dn > 0.$$

For $t \ll 1$ we have $a - a_0 \sim t^2$. $\epsilon + 3p = 0$ at the point of inflection of the $a(t)$ curve, and beyond $a \sim t^{1/2}$.

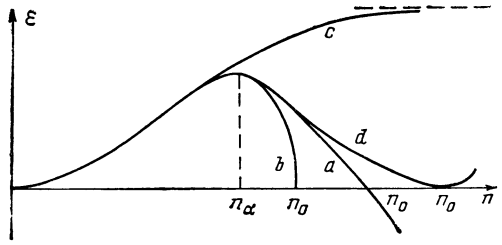
The solution for $a(t)$ can be symmetrically continued into the region $t < 0$:

$$a(t) = a(-t). \tag{19}$$

It is natural to assume that this symmetry extends to all physical properties, so that

$$\Psi(t) = \Psi^*(-t), \tag{20}$$

where Ψ is the state vector. The entropy of a certain part of the universe is $S(t) = S(-t)$, $S(0) = 0$. Such a "doubling" of physical reality is by virtue of its identical character not very meaningful, nor does it lead to any difficulties in principle. But on the other hand this eliminates the customarily raised question: "And what when $t < 0$?"



We shall consider in what follows also a variant of the equation of state (a plot of which is shown in the figure, curve b), for which $d\epsilon/dn \to \infty$ as $\epsilon \to 0$. If we assume in this limiting case that $\epsilon \sim (a - a_0)^{1/2}$, then we can readily obtain from (3) the dependence of $a - a_0$ on t (it is obviously logical to consider in this equation only the case C = 0):

$$a - a_0 \sim |t|^{4/3}. \tag{21}$$

In this case, as in the case of curve a, it is logical to assume symmetry of the states under time reversal. An alternate solution of the problem of negative time is possible (for C = 0), if the equation of state has the form shown on curve c ($\epsilon \to$ const as $n \to \infty$) or on curve d [$\epsilon \to A(a - a_0)^2$ as $a \to a_0$]. In case c, at the initial stage of the expansion of the universe, so long as $\epsilon \approx$ const, $p = -\epsilon$, $\ddot{a}/a$ = const > 0, i.e.,

$$a = e^{\lambda t} \quad \text{as } t \to -\infty. \tag{22}$$

Subsequently, when $\epsilon \neq$ const ($\epsilon \sim n^{4/3}$), the exponential growth is transformed into a growth like $a \sim t^{1/2}$. Analogously, for case d:

$$a - a_0 \sim e^{\lambda t} \quad \text{as } t \to -\infty.$$

We can attempt to determine the true character of the equation of state at $n \sim 1$ by comparing the observational data with the deductions of the theory concerning the magnitude of the initial density inhomogeneities.

5. QUANTUM THEORY OF OCCURRENCE OF INITIAL PERTURBATIONS AT $n \sim 1$.

Small density perturbations can be described with the aid of a wave function that depends on the dynamically-independent generalized "normal" coordinates z_K. In the case of statistical independence this function is of the form

$$\Psi(t, z_1, z_2 \ldots) = \prod_{\varkappa} \Psi_\varkappa(z_\varkappa, t). \tag{23}$$

Each of the functions Ψ_K [1] is described by its own Schrödinger equation of the harmonic-oscillator type (24), with the "mass" m(t) and "elasticity" r(t) as variables (it is not necessary here to have m > 0 and r > 0)

$$-\frac{1}{2m}\frac{\partial^2\Psi}{\partial z^2} + \frac{rz^2}{2}\Psi = -i\frac{\partial\Psi}{\partial t}. \tag{24}$$

Here

$$m = \frac{a^2}{\varkappa^2}\frac{d\varepsilon}{dn}, \tag{25}$$

$$r = -\frac{a^2}{\varkappa^2}\frac{d\varepsilon}{dn}\left\{4\pi n\left(\frac{d\varepsilon}{dn} + 3\frac{dp}{dn}\right) - 8\pi\frac{dp}{d\varepsilon}(\varepsilon + 3p) + 3\frac{\dot{a}}{a}\frac{d}{dt}\left(\frac{dp}{d\varepsilon}\right)\right\} + n\frac{d^2\varepsilon}{dn^2}. \tag{26}$$

The Schrödinger equation with a potential energy proportional to z^2 has two classes of self-similar solutions:

$$\Psi_+(z, t) = \nu_+(t)e^{-\mu(t)z^2}, \tag{27}$$

$$\Psi_-(z, t) = \nu_-(t)e^{-\mu(t)z^2}z. \tag{28}$$

These solutions satisfy (24), if the complex parameters μ, ν_+, and ν_- satisfy the following ordinary differential equations:

$$d\mu/dt = i(2\mu^2/m - r/2), \tag{29}$$

$$d\nu_+/dt = i\mu\nu_+/m, \tag{30}$$

$$d\nu_-/dt = 3i\mu\nu_-/m. \tag{31}$$

The initial values of ν and μ can be arbitrary (with Re μ > 0).

We note that the general solution of the Schrödinger equation can be represented in the form of a contour integral with respect to the parameter μ. For example,

$$\Psi(z, t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} dy_0 e^{-\mu(t)z^2}[C_+(y_0)\nu_+(t) + zC_-(y_0)\nu_-(t)]. \tag{32}$$

At the initial instant of time we have here $\mu(0) = x_0 + iy_0$, x_0 = const > 0, and the time-independent functions C_+ and C_- are determined from the integrals

$$\nu_+(0)C_+(y_0) = \int_{-\infty}^{+\infty} dz\,|z|\,e^{-\mu(0)z^2}\Psi(z, 0) \tag{33}$$

and analogously for C_-. A similar representation in the form of multiple integrals can be written for the elements of the Landau-Neumann density operator (the case of a "mixture").

We assume that each of the degrees of freedom of z_K is described by a wave function of the type (27)

$$\Psi_+ = \nu_+(t)e^{-\mu(t)z^2}.$$

We put below

[1] We shall henceforth omit the index κ in most cases.

$$\mu = x + iy = \varkappa^{-2}(X + iY). \tag{34}$$

Obviously

$$\overline{z^2} = 1/4x \tag{35}$$

or (in order of magnitude)

$$z \sim \varkappa X^{-1/2}. \tag{36}$$

We omit in (29) the small term $nd^2\epsilon/dn^2 = dp/dn$ from formula (26) for r [we confine ourselves to the case of "long" waves; this is possible, if the first term in $r \backsim (a/\kappa)^2 d\epsilon/dn$ is not anomalously small]. We find that the equation for $\kappa^2\mu = X + iY$ no longer contains κ. We arrive at the conclusion that if the solution of Eq. (29) exists at all, then the initial inhomogeneity $z_0(\kappa)$ should be proportional to κ and should not contain other parameters which differ from unity in order of magnitude. Thus, for long-wave perturbations

$$z \sim \varkappa/\dot{a}^2. \tag{37}$$

Using the law of averaging (17), we obtain

$$F(N) = \Delta N/N \sim N^{-5/6}/\dot{a}^2 \tag{38}$$

(with a proportionality factor ~ 1).

For $N \lesssim M^{-3}$ we need a correction for the pressure effect in the second stage:

$$F \sim N^{-5/6}B(N)/\dot{a}^2, \tag{38a}$$

where

$$[B(N)]^2 = \int_0^\infty B^2(\varkappa) z_0^2 \varkappa^2 e^{-\varkappa^2/4\alpha^2} d\varkappa \Big/ \int_0^\infty z_0^2 \varkappa^2 e^{-\varkappa^2/4\alpha^2} d\varkappa.$$

At the present time $\dot{a}^2 = 10^{-52}$—10^{-53}. Therefore, the formation of clusters with $F(N) \sim \Delta N/N \sim 1$ could have occurred up to the present time for masses containing $N = 10^{(6/5)52}$ or $10^{(6/5)53}$ particles, that is, masses of the order of $m = 3 \times 10^5$—$3 \times 10^6 M_\odot$. In order of magnitude this is the mass of a globular cluster.

The detailed discussion of the physical processes occurring when $n \sim 1$, contained in this paragraph, is of necessity beyond the level of our present-day knowledge. But even if our notions change very radically, the correctness of formula (38) is not excluded. A solution of (29) over the entire time interval undoubtedly exists if stability (that is, $r/m > 0$) corresponds to the initial instant of time $t \to 0$ or $t \to -\infty$. This criterion is satisfied by the equation of state represented by curve a of the figure ($r < 0$, $m < 0$), and is not satisfied by the equations of state of curves c and d. In these two cases, by virtue of the condition $d\epsilon/dn \to 0$ as $n \to n_0$ or ∞, the principal term in the "elastivity" r is the term of the spatial pressure gradient dp/dn (short-wave term), which has in both cases a sign opposite to that of $d\epsilon/dn \backsim m$. Therefore $dp/d\epsilon < 0$, which leads to a short-wave instability in the initial period of time as $t \to -\infty$. But even if an account of the nonlinear terms ensures that "initial" perturbations are finite (which is doubtful), there is no doubt that the amplitude of the initial perturbations obtained in such theories is too large (gigantic clusters of galaxies will be produced). For these reasons we turn to curves a and b.

For a unique determination of $\mu(0)$ in the case of curve a we must stipulate in addition

$$\mu(0) = 0. \tag{39}$$

In this case

$$\mu(0) = {}^1/_2(rm)^{1/2}, \quad z(0) = 1/2\sqrt{2}(rm)^{1/4}. \tag{40}$$

However, the requirement that the wave function (39) be stationary for the nonstationary state ($\ddot{a} > 0$) does not seem quite logical to us. It is therefore possibly of interest to consider the equation of state of curve b. In this case

$$a - a_0 \sim t^{4/3}, \quad m \sim -t^{2/3}\varkappa^{-2}, \quad r \sim t^{8/3}\varkappa^{-2} \tag{41}$$

[the sign of r is determined by summation of all three terms in the curly brackets of (26)]. Among the solutions of (29) there is one distinct solution that admits of a power-law approximation as $t \to 0$ (self-similar instability)

$$\mu = C_0 t^{-5/3}\varkappa^{-2}. \tag{42}$$

Here C_0 is a definite constant ~ 1, $\mathrm{Re}\, C_0 > 0$, and $\mathrm{Re}\, \mu \to \infty$ as $t \to 0$, that is, with such an equation of state the matter in the universe experiences in the initial state no density fluctuations.

We now consider the behavior of the solution of the system of equations for X and Y [equivalent to Eq. (29)] as $t \to \infty$ and as $t \to t_\alpha$ [t_α is the instant when $m = (a/\kappa)^2 d\epsilon/dn$ vanishes[2]]. Assuming that $\epsilon \to n^{4/3}$ as $t \to \infty$, we then have an asymptotic solution

$$Y = \frac{2}{3}\left(\frac{32\pi}{3}\right)^{1/4} t^{-1/2} \qquad X = B_\infty t^{-2}. \tag{43}$$

The constant B_∞ can be arbitrary if the inequality $X \ll Y$ is satisfied. This asymptotic solution describes an increase in z proportional to t, corresponding to the result of the classical (that is, not quantum) theory. If the constant B_∞ does not depend on κ, it is obvious that we have

[2] É. B. Gliner (in a paper now in press) calls states with $d\epsilon/dn = 0$, in connection with the isotropic character of the four-tensor $T_k^i \backsim \delta_k^i$, a μ-vacuum.

the already mentioned result (38), $\Delta N/N \sim N^{-5/6}\dot{a}^{-2}$. Near the point $t = t_\alpha$, the integral curve passes (through an anti node-type singular point).

$$Y = Y_\alpha + A_\alpha(t - t_\alpha)^2, \quad X = B_\alpha(t - t_\alpha)^2. \qquad (44)$$

When $t < t_\alpha$ it is obvious that A_α and $B_\alpha \sim 1$, since $X_0 \sim 1$ as $t \to 0$ in the case of an equation represented by curve a, or $C_0 \sim 1$ in the case of an equation represented by curve b. If analytic continuation of the solution in the vicinity of the point t_α is possible, then A_α and B_α are continuous and $B_\infty \sim 1$. However, the possibility of analytic continuation gives rise to certain doubts, for when $t \to t_\alpha$ we have $z \sim 1/(t - t_\alpha)$, that is, $z \to \infty$ (according to the linear theory). Actually it is necessary to take account near the point t_α of nonlinear effects.

The singular character of the point t_α can be greatly reduced by going over to a reference frame with $t_1 = 0$. This causes us to assume that a change in the order of magnitude of z at the point t_α can hardly take place [although it is quite possible that when $t > t_\alpha$ the statistical state of each elementary "oscillator" z_K is no longer described by the wave function (27), but by a density operator].

6. COSMOLOGICAL HYPOTHESIS

Let us assume that formula (38a) is correct. In this case the instant of formation of clusters containing N particles with mass $m = NM$ is determined by the condition $F(N, t) = \Delta N/N \sim 1$, that is, (taking into account that $\dot{a} \sim t^{-1/3}M^{1/3}$)

$$t(N) \sim N^{5/4}M[B(N)]^{-3/2} \sim 10^2 \text{ years} \cdot \left[\frac{m}{M_\odot}\right]^{5/4} B^{-3/2}. \qquad (45)$$

The first to be formed are obviously "primordial" stars with mass $m \leq 0.4M_\odot$ (at $t \sim 10^2$ years). From similarity considerations we assume that the fraction of the primordial gas not captured by the stars decreases in accordance with a power law of the form $[t(N_\odot)/t]^\alpha$. Under likely values of the exponent α, no more than several per cent gas remain at a time corresponding to 10^9 years.

How do we explain the formation of galaxies [for which, according to (45), $t(N) \sim 10^{16}$ years]? The answer presented here for discussion contains several hypothetical assumptions.

Clusters containing not more than a certain critical number of primordial stars will sooner or later experience the gravitational collapse of Tolman-Oppenheimer-Snyder-Volkoff (the question of the evolution of stellar clusters, leading to collapse, has recently been discussed in connection with the problem of quasars; see, for example,[4]; the duration of the evolution decreases if a gaseous or dust-like phase is present in the cluster).

An estimate shows that even after $t \sim 10^6$ years it is possible for quasars with a mass of 500 $M_\odot$, for which the sum of the formation time given by (45) and the time of hydrogen burn-up are close to minimum and of the order of 10^6 years, to collapse. This is followed by collapses of larger clusters of matter. According to our hypothesis, which is apparently confirmed by the data on quasar observation, several per cent of the matter is ejected during the collapses, with jet velocities of the order of 10^3—10^4 km/sec. (The ejection of gas clouds from M-82 confirms the lower of these figures.) The collapse results in a "post-collapse" object (PC-object), which has very small dimensions and is manifest principally through its gravitational field.

It is possible that the bulk of the matter in the universe is contained at present in PC-objects; with this assumption, an estimate of the average density of matter, in which only galaxies are taken into account, turns out to yield much too low a value and the disparity with the condition $C = 0$ disappears (it is precisely in this context that several authors have discussed this hypothesis, see, for example, the article by Zel'dovich[1] with a reference to I. D. Novikov). Incidentally, the intergalactic primordial stars with low luminosity, and intergalactic gas and dust of low density, are also very difficult to observe, and we do not exclude the possibility that it is just these components that constitute a principal or appreciable fraction.

We propose that the process of formation of galaxies begins at $t \sim 10^9$ years and still continues, and will continue to infinity, encompassing ever increasing masses; the non-uniformity of the gas distribution is first produced in this case by jets due to collapses of clusters of primordial stars and during the later stages an ever increasing role will be assumed by collapses of cores of galaxies and clusters of PC-objects. By the time $t \sim 10^9$ years is reached the proposed fraction of the gas component is approximately 10 per cent; these are remnants of the primordial gas and of gas produced by small collapses. At $t \sim 10^9$ years there occur collapses of clusters containing 10^5M . The kinetic energy of the jets ejected in such collapses can reach, according to our hypotheses, 10^{53}—10^{55} erg; this energy goes over

into the energy of the shock waves propagating through the gas. Streams of gas are produced, and in regions where several streams meet the density of the gas increases by a factor of several times. As a result of the gravitational contraction of these condensations, galaxies are subsequently produced (possibly in pairs or in groups), with masses amounting to approximately $\frac{1}{3}$—$\frac{1}{5}$ of the mass of the gas. If the time for the formation of the condensations is 10^9 years, then the formation of a cluster with mass 10^{44} g requires a velocity of 10^6 cm/sec, that is, an energy of 10^{56} erg. According to our assumption, this is the energy of 10^1—10^3 collapses of masses equal to $10^5 M_\odot$, or 1—10^2 collapses of masses of $10^6 M_\odot$.

The gas clouds which become condensed in collisions of jets should possess in most cases an angular momentum and constitute the flat component of the galaxies. To complete the picture, we must assume, that after the condensed regions of gas have been formed, some small fraction (~ 1 per cent) of the globular clusters of primordial stars "left intact" by the collapses, and a similar fraction of the PC-objects, is captured by the galaxies, thus forming the spherical components of galaxies. In accordance with the observations, the spherical component of a galaxy practically does not participate in the galactic rotation of the gas and in the formation of stars of secondary origin from the gas.

The mass of single PC-objects (that is, not included in the clusters and in the core) in our galaxy is apparently much smaller than the mass of the galaxy, otherwise their gravitational field would be manifest in the speed of galactic rotation. (This does not exclude a considerable number of PC-satellites of the galaxy with former semi-axes.)

Globular clusters not captured by galaxies have not been observed. This may be due to their lack of bright and variable stars, characteristic of galactic clusters, which apparently arise (or are captured) when galactic clusters pass through the core of the galaxy.

Of course, the picture described is very hypothetical and must be supplemented, refined, or rejected as a result of further considerations.

The author is grateful to Ya. B. Zel'dovich, numerous discussions with whom led to formulation of the entire problem as a whole and enriched the work with many ideas. The author is also grateful to I. E. Tamm and E. M. Lifshitz for a discussion and valuable remarks.

[1] V. A. Fock, UFN 80, 353 (1963); Ya. B. Zel'dovich, UFN 80, 357 (1963); E. M. Lifshitz and I. M. Khalatnikov, UFN 80, 391 (1963); Translations: Soviet Phys. Uspekhi 6, 473 (1964), 6 475 (1964), and 6, 495 (1964).

[2] E. M. Lifshitz, JETP 16, 987 (1946).

[3] B. Bok and P. Bok, Mlechnyĭ put' (The Milky Way), Fizmatgiz, 1959, p. 117.

[4] Ya. B. Zel'dovich and I. D. Novikov, UFN 84, 377 (1964) and 86, 447 (1965), Soviet Phys. Uspekhi 7, 763 (1965) and in press.

Translated by J. G. Adashko
45