Gravitational Instability: 
An Approximate Theory for Large Density Perturbations 

Y.A. B. Zeldovich
Institute of Applied Mathematics, Moscow

Received September 19, 1969

An approximate solution is given for the problem of the growth of perturbations during the expansion of matter without pressure. The solution is qualitatively correct even when the perturbations are not small. Infinite density is first obtained on disc-like surfaces by unilateral compression.

The following layers are compressed first adiabatically and then by a shock wave. Physical conditions in the compressed matter are analysed.

Key words: Galaxies formation — Cosmology — Gravitational instability

1. The Approximate Solution

The linear theory of perturbations, applied to the uniform isotropic cosmological solution, is now well understood. It is generally admitted that its predictions are limited by $\delta \rho / \rho < 1$, and that further events must be followed by numerical calculations. Such calculations, in three dimensions and with random initial conditions, promise to be tedious. Therefore an approximate method, which gives the right answer at least qualitatively, is of interest.

In this article the linear theory is taken to formulate the answer in terms of lagrangian coordinates: the actual position $\mathbf{r}$ of a particle is given as a function of its lagrangian coordinate $\mathbf{q}$ (i.e., its initial position) and the time $t$, $\mathbf{r} = \mathbf{r}(t, \mathbf{q})$. The linear theory is applied to the simplest case of pressure $\rho = 0$ ("dust") in the Newtonian approximation. Only the growing perturbations are considered. The answer is of the form

$$ \mathbf{r} = a(t) \mathbf{q} + b(t) \mathbf{p}(\mathbf{q}). $$

(1)

The first term $a(t) \mathbf{q}$ describes the cosmological expansion; the second term describes the perturbations. The functions $a(t)$ and $b(t)$ are known; $b(t)$ is growing more rapidly than $a(t)$, as a result of gravitational instability. The vector function $\mathbf{p}(\mathbf{q})$ depends on the initial perturbation. With given $\mathbf{r}(t, \mathbf{q})$, it is possible to calculate the distribution of velocity and density in space; $\mathbf{r}(t, \mathbf{q})$ contains the whole picture of the motion.

The approximation proposed in this article consists in the extrapolation of formula (1) into the region where the perturbations of density $\delta \rho / \rho$ are not small.

Let us first investigate the consequences of the approximation; this will help us to analyse its plausibility. In order to follow the behaviour of a small group of particles centered on some definite $\mathbf{q}$, we calculate the tensor of deformation

$$ D_{ik} = \frac{\partial r_i}{\partial q_k} = a(t) \delta_{ik} + b(t) \frac{\partial p_i}{\partial q_k}. $$

The derivatives $\frac{\partial p_i}{\partial q_k}$ define a set of fundamental axes.

After choosing the coordinate system along the axes, one obtains1) for a given $\mathbf{q}$

$$ D = \begin{pmatrix} a(t) - \alpha b(t) & 0 & 0 \\ 0 & a(t) - \beta b(t) & 0 \\ 0 & 0 & a(t) - \gamma b(t) \end{pmatrix}. $$

A volume which was initially a cube (at $t \to 0$) and which would be a cube in the unperturbed motion, is transformed into a parallelepiped. One can always choose the axis of the cube so that it is transformed into a rectangular parallelepiped; the axes are not rotating in solution (1). The density near a particle with given $\mathbf{q}$ is given by the conservation of mass

$$ q(a - \alpha b) (a - \beta b) (a - \gamma b) = \bar{\rho} a^3. $$

(2)
We recall that $\alpha$, $\beta$, and $\gamma$ are functions of the point $q$; $a(t)$ and $b(t)$ are the same for all particles. If $\alpha(q) > 0$, one can specify the moment when $q \to \infty$ by $a(t) - \alpha b(t) = 0$. In a given $q$ volume, we find the point where $\alpha$ has its highest value $\alpha_m$; this locates the particle in which the density first goes into infinity, at some time $t_\alpha$, for which $a(t_\alpha) - \alpha_m b(t_\alpha) = 0$.

The most important point to be emphasized is that infinite density results from uniaxial compression in the direction of the $x$-axis. The probability of the coincidence of $x$ and $\beta$, or of a triple coincidence $\alpha = \beta = \gamma$, is zero.

The picture is very different from a spherically symmetric (SS) compression. The SS case was considered due to its simplicity; I think that it is degenerate and not typical of the general case of random initial perturbations. Later, at $t > t_\alpha$, formula (2) is not applicable to particles which have gone through $q = \infty$; the matter stays compressed. But we feel that it is still possible to apply (1) and (2) to other particles. By continuity, particles with high $\alpha$, $\alpha_m - \alpha_i < \alpha_m$, surround the "maximal" particle, lying on some triaxial ellipsoid. The direction of the fundamental axis of $\frac{\partial F}{\partial q}$ also varies slowly, so that the $\alpha$ direction is nearly the same as long as $\alpha_i$ is near $\alpha_m$.

The uniaxial compression makes the three-dimensional ellipsoid in $q$-space into a flat two-dimensional ellipse in the real r-space. The volume density $q$ is infinite, but the product of $q$ times the width $l$ (equal to the density per unit of surface $a$) is finite on this ellipse.

2. Comparison of Various Approximations

Let us return to the motivation of (1), on which the picture proposed is based.

In the linear approximation, there are other formulations, for example

\[ \rho(r, t) = \bar{\rho}(t) \left( 1 + f(t) \delta \left( \frac{r}{a(t)} \right) \right), \]  

with known $f(t)$, $a(t)$ and $\delta$ given by the initial conditions. Why should one prefer (1) to (3)? The proposed solution (1) always gives finite velocity and finite acceleration for a particle, up to the moment when this particle is splashed against other particles, giving $q = \infty$. Matter with infinite density $q$ forms discs with finite density $\sigma$. But by the properties of gravitational potential, $q = \infty$, $\sigma \to \infty$ gives a finite potential and a finite acceleration. The approximation proposed is exact at one extreme, when the perturbations are small. But the approximation gives only a finite error at the other extreme, when $q = \infty$.

This is in contrast with (3): if one attempts to extrapolate (3), meaningless negative densities are predicted for some parts of the volume $\delta < - \frac{1}{t}$, while in other parts the density is only doubled.

The better performance of (1) as compared with (3) can be traced back to the fundamental equations. The solution (3) corresponds to equations written in eulerian form. During linearisation, terms $\nabla \cdot (v \delta)$ and $\left( \nabla \nabla \right) v$ are neglected, where $\delta = \frac{\partial q}{\partial q}$, and $v = u - Hr$ is the peculiar velocity.

By adopting solution (1), we adopt a law of motion for every particle:

\[ u = \dot{\alpha} q + \dot{b} p(q); \quad \frac{du}{dt} = \dot{\alpha} q + \dot{b} p(q). \quad (4) \]

Given formula (1), the density is calculated exactly.

The only error is in the use of the perturbation of gravitational force $\delta F$, acting on a particle $q$, where $\langle \delta F \rangle$ is linearised as a function of the perturbation of position of the particle considered $b p(q)$, and of all other particles with differing $b p(q')$. The analytic evaluation of the error is extremely difficult. Djachenko proposes to make a numerical estimate of the error, by taking solution (1) with a definite $p(q)$ at a definite moment $t$, calculating the actual distribution of $\rho(r, t)$, the gravitational potential $\phi(r, t)$ and the force $F(r, t)$, and comparing this force with the acceleration given by (4) for the approximate solution. As long as this trial is not made, the real accuracy of the solution for perturbations of different amplitude is unknown. Still, the main qualitative conclusion about the uniaxial type of compression seems to be inescapable.

The compression is due to gravitational interaction: it is due to gravitation that $b(t)$ grows faster than $a(t)$ in (1). The initial excess of density near some particle $q_0$ increases because other particles are attracted to $q_0$. But solution (1) and expression (2) for the density also take into account the tidal forces from neighboring perturbations which destroy the spherical symmetry of compression. At small $t$ and small $\delta q/q$ formula (2) coincides with (3):

\[ \rho(q, t) = \bar{\rho} \left( 1 + (\alpha + \beta + \gamma) \frac{b(t)}{a(t)} \right) = \bar{\rho} (1 + S(q) l(t)). \quad (5) \]

The initial growth of the perturbation depends only on the sum $\alpha + \beta + \gamma = S(q)$. But later, as shown by (2), all three parameters $\alpha$, $\beta$ and $\gamma$ are of importance. Therefore (1) and (2) contain more information than (3). The later non-linear behaviour of the density
is not uniquely determined by the initial density amplitude in the linear period; it also depends on the spatial distribution of velocity and density in neighbouring regions, on which \( \alpha, \beta, \) and \( \gamma \) depend. This important point is overlooked if one takes the spherically symmetric case as a model for the nonlinear situation.

3. Astronomical Implications

Assuming that the approximate solution is qualitatively true, we must discuss a) whether the necessary conditions (\( \mathcal{O} = 0 \), newtonian approximation) are fulfilled, b) what physical processes occur in the compressed regions, and c) the place of the solution in the problem of the structure of the Universe. The answer to a) depends on the type of initial perturbation. Widely discussed are adiabatic perturbations, characterised by \( \frac{\delta T}{T} = \frac{1}{3} \frac{\delta \rho}{\rho} \) before recombination.

As shown by Silk (1967), the photon viscosity eliminates perturbations of small scale \( (\mathcal{M} < 10^{12} \, M_\odot) \). On the other hand, it is plausible that the spectrum is decreasing for greater scale. Therefore the perturbations with \( M \sim 10^{12} \, M_\odot \) are the most important.

We apply the approximate theory to the period after recombination, assuming that the perturbations are small at \( Z = 1400 \) (just after recombination) and grow so that galaxies, etc., are formed before \( Z = 0 \) (the present day).

After recombination, the Jeans’ mass (depending on the neutral gas pressure) is of the order of \( M_J = 10^5 - 10^6 \, M_\odot \) (Doroshkevich, Zeldovich and Novikov, 1967; Peebles and Dicke, 1968). The situation \( M \approx 10^{12} \, M_\odot > M_J \) means that pressure can safely be neglected. At the moment of recombination the event horizon (the sphere with \( r = ct \)) contains \( M_h = 10^{12} \, M_\odot \), so that \( M < M_h \). Therefore newtonian theory is applicable. The inequalities are even better fulfilled later on at \( Z < 1400 \), because \( M_J \) diminishes and \( M_h \) grows during further expansion.

b) Now we turn from the premises to the consequences of the approximate solution. To study the characteristic unilateral compression, we neglect the motion in other directions and consider the unidimensional problem. The subsequent evolution is shown in three figures: 1 - for small perturbations \( (t < t_c) \), 2 - for the moment when \( q \) just attains infinity on the line considered \( (t = t_c) \), and 3 - for \( t > t_c \). On each figure the curve \( r = f(q) \) is given. To be precise, it is \( r_p \) as a function of \( q_\perp \), the \( z \)-axis being chosen in the direction of maximal deformation.

The quantity of matter between a pair of points is proportional to \( q_\perp - q_\parallel \), because \( q \) is the lagrangian coordinate. In the physical \( (r) \) space it is contained in the strip \( r_p - r_\parallel \); therefore the density \( \rho \) is proportional to \( \frac{q_\perp - q_\parallel}{r_p - r_\parallel} \cdots \). This is the unidimensional simplification of formula (2). In Fig. 2 the curve has a point with a horizontal tangent. When such a point occurs for the first time, we must have \( \frac{dr_r}{dq} = 0 \), \( \frac{d^2r}{dq^2} = 0 \) at \( t = t_c \); \( q = q_c \).

Expanding \( r(q) \) near \( q_c \) it is easy to obtain

\[
q \sim \left( \frac{dr_r}{dq^2} \right) (q - q_c)^{-2}, \quad r = r_c + \frac{1}{6} \left( \frac{d^2r}{dq^2} \right) (q - q_c)^3.
\]

Thus \( q \sim (r - r_c)^{-2/3} \) when \( t = t_c \).

Going to Fig. 3 one would obtain \( q = \infty \) at points of maximum \( (q_\parallel) \) and minimum \( (q_\perp) \) by formal application of the formulae. But the particle \( q_\parallel \) cannot reach the \( r_{\text{max}} \) shown on the curve; to do this, it would have to jump over the particle \( q_c \). Suppose that the density of the disc is infinite; then all particles reaching \( r = r_c \) will abruptly come to a standstill by encounter with the disc. But this produces a receding shock-wave. In the encounter, the kinetic energy of relative motion is transformed into heat. The matter is no longer cold; its density is just 4 times the density before the shock. As \( t - t_c \) grows, the velocity of impact grows \( \sim \sqrt{t - t_c} \), but the density of matter going into the shock decreases as \( (t - t_c)^{-1} \). The pressure \( \sim q u^2 \) remains constant in the first approximation (this means that there is no power dependence on \( t - t_c \)).

Some physical quantities are evaluated below for the case \( \mathcal{O} = 1 \). The linear scale of the perturbations is given by the corresponding characteristic mass \( M \), and the amplitude of the perturbations by the moment \( t_c \) or the corresponding redshift \( Z_c \). Only orders of magnitude are given; detailed calculations are postponed to a comprehensive paper to be published in “Astrophysica” (edited by Ereven, USSR).

The surface matter density \( \sigma \) is given by

\[
\sigma = 10^{-4} \left( \frac{M}{10^{12} \, M_\odot} \right)^{1/3} \left( 1 + Z_c \right)^{3/2} \left( Z_c - Z \right)^{1/2}.
\]

The impact velocity \( v \) (cm/s) is

\[
v = 10^7 \left( M/10^{12} \, M_\odot \right)^{1/3} \sqrt{Z_c - Z}.
\]

The pressure in the compressed matter is given by

\[
\mathcal{P} = 2.5 \cdot 10^{-16} \left( 1 + Z_c \right)^4 \left( M/10^{12} \, M_\odot \right)^{3/2}.
\]
The maximum density (in the adiabatically compressed matter) is

$$\rho_{\text{max}} = 4 \cdot 10^{-25} (1 + Z_{\odot})^{3/5} (M/10^{12} M_\odot)^{3/5}.$$  

Minimum density corresponds to the pressure given above and a temperature of the order of 5000° — at higher temperatures the radiation energy losses are great and the temperature drops. This gives

$$\rho_{\text{min}} \approx 5 \cdot 10^{-28} (1 + Z_{\odot})^4 (M/10^{12} M_\odot)^{1/3}.$$  

For

$$M = 10^{13} M_\odot, Z_{\odot} = 4 \text{ and } t_c = 1.5 \cdot 10^{14}\text{ s},$$

we have

$$\rho_{\text{max}} = 2 \cdot 10^{-25}, \quad \rho_{\text{min}} = 3 \cdot 10^{-25}.$$  

Evaluating the formation of a star with the Schmidt formula (Schmidt, 1959) one obtains a negative result, the time needed being greater than $t_c$.

The stars are formed at a later stage and not simultaneously with the disc.

c) In the most favourable case, the approximate solution describes a definite part of the evolution. But it is certainly not intended to cover the whole theory of formation of galaxies. The discs are not in equilibrium. The deformation in the plane of the discs, given by the parameters $\beta$ and $\gamma$ (2), occurs in both directions. Nearly half of the matter, compressed in the discs, is not gravitationally bound.

It seems that the formation of discs is an unavoidable result of a definite set of assumptions about the initial perturbations. But even if this hypothesis is confirmed, it is difficult to predict how much of the discs would remain in the structure of near-by contemporary galaxies.

Another problem as yet unsolved, is the possible application of the hypothesis to fluctuations of entropy, i.e., fluctuations of matter density at constant temperature. It is assumed that the spectrum is decreasing, so that after recombination the most
important fluctuations are those on the scale of the Jeans’ mass \((10^8 \pm 10^9 M_\odot)\). On this scale the gas pressure is important and the approximate solution is inapplicable. Pressure works against the unilateral compression; the formation of protostars (Doroshkevich, Zeldovich and Novikov, 1967) or globular clusters (Peebles and Dicke, 1968) is plausible. But nevertheless the fluctuations of greater scale remain, and for them the gas pressure is negligible.

On a greater scale the fluctuations are smaller; therefore they lead to condensation later, after the globular clusters are formed. The question is, what part of the matter has gone into globular clusters? Do they now heat the remaining gas? And finally: should one apply the approximate solution to the “gas” whose atoms are globular clusters or protostars or small gas clouds?

4. Mathematical Appendix

Explicit formulae can be given for the functions \(a(t)\) and \(b(t)\) of (1). It is convenient to take as independent variable the redshift \(Z\) instead of the time. These are connected by

\[
t = \frac{1}{H_0} \int_0^\infty \frac{dZ}{(1+Z)^\frac{1}{2}(1+\Omega Z)},
\]

where \(H_0\) is the present value of the Hubble parameter \(\sim 100 \text{ km/s megaparsec.}\)

It is easy to make the integration

\[
H_0 t = \frac{\sqrt{1+\Omega Z}}{(1-\Omega)(1+Z)} \frac{\Omega}{2(1-\Omega)^{3/2}} \ln \left(\frac{\sqrt{1+\Omega Z}+\sqrt{1-\Omega}}{\sqrt{1+\Omega Z}-\sqrt{1-\Omega}}\right),
\]

but to extract the limiting cases (\(\Omega = 0\), or \(\Omega = 1\), or \(\Omega < 1\), \(\Omega Z > 1\) etc.) it is better to use the integral directly.

The function \(a(t)\) (see (1)) is replaced by

\[
a(Z) = \frac{1}{1+Z}.
\]

The lagrange variable \(q\) is defined by (A.3) so that it coincides with \(r\) in the case of unperturbed motion at the present time \((Z = 0)\). The growth of perturbations is given by

\[
b(Z) = \sqrt{1+\Omega Z} \int_Z^\infty \frac{dZ}{(1+Z)^{3/2}(1+\Omega Z)^{3/2}}.
\]

We recall that \(Z\) decreases with increasing time. For \(\Omega = 1\) (flat universe)

\[
t = 2/3 H_0^{-1}(1+Z)^{-3/2}; \quad a(Z) = 1/1 + Z = (3/2 H_0 t)^{3/2}
\]

and

\[
b(Z) = 2/5 (1+Z)^{-2} = 2/5 (3/2 H_0 t)^{4/3}.
\]

For \(\Omega \ll 1\), the growth of perturbations effectively stops at \(Z \sim 1/\Omega\); for the ratio \(b/Z/a(t)\) or \(b(Z)/a(Z)\) increases more slowly than a power of \((1+Z)^{-1}\) when \(Z\) decreases after \(Z = 1/\Omega\). For the sake of completeness, the second (damped) mode of perturbations is given:

\[
b_2(Z) = \sqrt{1+\Omega Z} \frac{\partial a}{\partial t} = \frac{\partial a}{\partial Z} \frac{dt}{dZ}.
\]

The perturbations are given in terms of displacements \((r)\), but not the usual \(\delta q;\) from (2), it is clear that \(\delta q \sim b/a \sim f(t)\) in the linear stage. All expressions given above are written for matter without pressure after recombination, \(Z < 1400\). If \(\Omega \leq 0.03\), the density of radiation remains greater than the matter density for some time after recombination and the gravitational action of radiation changes the expansion law \(d(t)\) or \(a(Z)\) and the perturbation law \(b(Z); b_2(Z)\) is also changed in the interval \(1400 > Z > 40000 \Omega\).

Suppose that at the moment of recombination, \(Z = Z_r = 1400\), the perturbations are given by \(\delta q = \delta (r)\), \(\nu = \nu (r)\). They are assumed to be small; for simplicity the case \(\Omega > 0.03\) is considered.

To obtain the function \(p(q)\) of (1), it is advisable to work with Fourier decomposition. First we transform from \(r\) to \(q\) by \(r = a q = 1 + Z\) (here at \(Z_r\) perturbations can be neglected).

Having constructed \(\delta (q)\) and \(\nu (q)\) (all for \(t = t_r\), \(Z = Z_r\) at the moment of recombination), we decompose them:

\[
\delta (q) = \int \delta (k) e^{i q \cdot k} d^3 k,
\]

\[
\nu (q) = \int (n v_1 (k) + l_2 v'_1 (k) + l_3 v''_1 (k)) e^{i q \cdot k} d^3 k,
\]

where

\[
n = \frac{k}{|k|}, \quad l_1 \perp n, \quad l_2 \perp l_1 \text{ and } l_2 \perp n.
\]

\(v_1\) is the longitudinal component and \(v'_1\) and \(v''_1\) the two transverse components of the velocity. The growing perturbation is given at \(Z_r\) by Doroshkevich and
Zeldovich (1963) as:
\[
\delta_x(k) = \frac{3}{5} \left( \delta(k) + i |k| v_1(k) (1 + Z_r) \right),
\]
\[
v_x(k) = n \frac{3}{5} \left( v_1(k) + \frac{i \delta(k)}{|k|(1 + Z_r)} \right).
\] (A.8)

The displacement corresponding to the growing mode (taking into account the fact that \( \Omega Z_r \gg 1 \)) is given by
\[
r - aq = \int v(q, t) dt = \frac{3}{4} v_1(q, t) t = b p(q). \quad (A.9)
\]

Because \( v_x \sim t^{\beta/3} \) and applying (1) to \( t = t_r \), we obtain
\[
p(q) = \int \frac{p(k)}{n v_1(k)} e^{i q \cdot k} d^3 k,
\] (A.10)

with
\[
p(k) = \frac{t_r}{5 \Omega(Z_r)} \frac{9}{20} n v_1(k) + \frac{i \delta(k)}{|k|(1 + Z_r)} = n \eta(k).
\]

The \( v_1(k) \) and \( \delta(k) \) without indexes are taken at \( t = t_r \).

After neglecting the transverse displacement (even if \( v_1 \sim v'_1 \sim v'_2 \) at \( t_r \), thereafter \( v_1 \) grows while \( v'_2 \) and \( v'_1 \) decrease), \( p(q) \) is vortex-free:
\[
\frac{\partial p_x}{\partial q_x} = \int n v_1(k) \eta(k) e^{i q \cdot k} d^3 k = \frac{k_x}{k_0} \eta(k)
\]
\[
\cdot \eta(k) e^{i q \cdot k} d^3 k = \frac{\partial p_x}{\partial q_x}.
\] (A.11)

Therefore \( p(q) \) can be written as the gradient of a scalar function \( \xi(q) \):
\[
p(q) = \nabla_{\xi} \xi(q),
\] (A.12)

with the Fourier image \( \xi(q) = \frac{1}{|k|} \eta(k) \).

It can be shown that the peculiar velocity given by the approximate solution is also vortex-free in physical space, as a function of the eulerian \( r \) co-

ordinates. So is the Hubble velocity, and therefore the total velocity also. The exact solution has this property because the motion occurs under the action of gravity — a force with potential. The approximate solution has this property even when the perturbations are not small \( (r - aq \) not neglected). This is one more argument in favour of the approximate solution. With a gaussian probability distribution of \( \delta(q), v(q) \), and the Fourier-components of \( \delta, v \) and \( \psi \), the distribution of \( \alpha, \beta \) and \( \gamma \) defined by (2) is not gaussian. As found by Doroshkevich,
\[
W(\alpha, \beta, \gamma) \sim (\alpha - \beta) (\alpha - \gamma) (\beta - \gamma)
\]
\[
\cdot \exp \left\{ - m^2 [\alpha^2 + \beta^2 + \gamma^2 - \frac{1}{2} (\alpha \beta + \alpha \gamma + \beta \gamma)] \right\}. \quad (A.13)
\]

The probability that all of them are positive \( (x > \beta > \gamma > 0) \) is 8\%, and the probability that \( x > \beta > 0 > \gamma \) is 42\%; by symmetry, there is a 42\% probability that \( x > 0 > \beta > \gamma \), and 8\% that \( 0 > x > \beta > \gamma \).

Gratitude is due to V.F. Djachenko and A.G. Doroshkevich for valuable ideas and discussion and to V. Chechetkin for help.

References


Ya. B. Zeldovich

Institute of Applied Mathematics

USSR Academy of Sciences

Miusekaja plozsad 4, Moscow A-47, USSR