On the Nature of Mass

F. HOYLE & J. V. NARLIKAR
Institute of Theoretical Astronomy, University of Cambridge

The increasing number of observations of discrepant redshifts means that no longer can these be passed off as chance juxtapositions. A possible explanation of the data is given here in terms of a theory that incorporates a gravitational "constant" that is decreasing with time.

EVIDENCE concerning "discrepant redshifts" has accumulated fairly rapidly in recent months. The case where it is hardest to deny the evidence was reported at the beginning of the present year by Arp[1]. The galaxy NGC 7603 was found to be connected to an appendage by two arms which intersect at the appendage $z=0.056$. In the customary Doppler interpretation of the redshift this difference exceeds 8,000 km s^{-1}. A redshift difference of the same order was found many years ago[2] by Minkowski, Humason and Zwicky in "Zwicky's triple system". The galaxies in this system are also connected by a bridge.

If it is accepted that a discrepancy of $\sim$8,000 km s^{-1} exists for NGC 7603, then there seems no good reason for denying the two cases with discrepancies $\sim$20,000 km s^{-1} obtained by Sargent (ref. 3, and W. L. W. Sargent's contribution to the Vatican Conference, 1970). Both the latter cases are compact groups of galaxies in which one member has the discrepant redshift. Less than ten such groups have been examined. The probability of a more distant galaxy being projected against a nearer cluster by chance is small. Arp (private communication) has obtained what appears to be a jet connecting the quasi-stellar object Makarian 205, $z \simeq 0.07$, to the galaxy NGC 4319. The discrepancy is again of order 20,000 km s^{-1}.

Three cases have been observed of a quasi-stellar object apparently associated with a small compact cluster of faint galaxies. In one case Gunn has reported[4] the same redshift for the QSO as for one of the galaxies of the cluster. In the other two cases there appear to be large redshift differences (Hazard, Jauncey, Sargent and Gunn, private communication).

In a recent preprint Burbidge, Burbidge, Solomon and Strittmatter have drawn attention to a bridge which appears on the Palomar sky prints to connect the quasi-stellar object PHL 1226 to the bright galaxy IC 1746. The redshift of PGL 1226 is $z=0.404$, whereas that of the galaxy has not yet been measured, but must be small because of its brightness.

Because all objects are projected on the sky there must be some apparently peculiar juxtapositions of objects that are really at very different distances. This has hitherto led to a situation in which all but a very few astronomers have dismissed apparent redshift discrepancies as simply unusual projection effects. Yet there has to be a point of balance in one's judgment as to the extent of the chance juxtapositions one is willing to accept. One of us was aware at an early stage of the case of Zwicky's system (Zwicky, private communication) but it seemed that, in spite of the bridge, there might be a very peculiar chance juxtaposition. But the case of VV 172[3] formed the point of balance and NGC 7603 has turned the balance.

If a highly convincing theory of discrepant redshifts were available then we think there is little doubt that the data would today be considered reasonably clear-cut. It has been the absence of such a theory that has caused most astronomers to prefer to believe in unusual projection effects. If, as seems very possible, the accumulation of data forces us over a watershed (not only in our thinking but in the history of astronomy) it will clearly become necessary to arrive at a theory of discrepant redshifts. We wish to emphasize the need for a thoroughly radical assessment of the problem, considering it unlikely that a satisfactory theory will be achieved by a small change in our concepts. Explicitly, we do not think discrepant redshifts will be explained adequately either as simple Doppler peculiar motions or as excess reddening due to gravitation.

In the following discussion we shall present the usual Friedmann cosmological models and the Einstein de Sitter case in particular, from an unusual point of view. Mathematically, this point of view is at first entirely equivalent to the usual formulation. Then at a later stage we shall arrive at a possible major shift from the usual theory. The balance between the usual theory and an entirely different view both of physics and astrophysics hinges on a question raised by Dirac[5] in 1937.

It is well known that when c and $\hbar$ are set equal to unity, only a single dimensionality is needed for the whole of physics. We take this to be length and we denote the unit by L. Every quantity has a dimensional form L^n, for example, mass $\sim L^{-1}$, frequency $\sim L^{-1}$, charge $\sim L^0$, magnetic field $\sim L^{-2}$ and the gravitational 'constant' has dimensionality L^2.

Every observation is concerned with a dimensionless number, so that every observation is concerned with a product of quantities such that the dimensional dependencies on L cancel to zero. So far as experimental physics and engineering are concerned it is possible to convert a quantity of dimensionality L^n into a quantity of dimensionality L^m by means of a linear device provided $n=m$. Non-linear devices are needed if $n \neq m$. This property makes it comparatively easy to see what kind of physical device is needed to relate one quantity to another. There is no doubt that physics and engineering are made quite unnecessarily complicated by the current practice of using multiunit systems.

Because observed quantities are dimensionless they are unaffected by a scale change in the length unit. But what if we elect to change the length unit in accordance with a well behaved function $\Omega(x)$ (no zeros of Ω as well as no infinities) which varies with the position of the space-time point X? Nothing should be changed provided that first, we always combine physical quantities at the same X, and, second, every physical quantity is scaled according to its dimensionality —a quantity of dimensionality L^n being scaled by Ω^n.

Physical theories with these properties are said to be conformally invariant. This is a different kind of invariance from the coordinate invariance of relativity. Not all physical theories, however, are conformally invariant. Maxwell's theory is conformally invariant. The derivative terms of Dirac's equation are conformally invariant, but the mass term is not. Einstein's theory is not conformally invariant and physical theories cease to be conformally invariant whenever "mass" is involved. This is because the second property is then not satisfied. Clearly, we cannot expect classical dynamics to be conformally invariant unless the classical action

$$S = -\int m\, da \tag{1}$$

of particle a is invariant. Since the element da of the path of the particle becomes multiplied by Ω we would require the mass m to be multiplied by Ω^{-1}, which indeed would accord with the dimensionality of mass, $\sim L^{-1}$. But Ω can now be different at different space–time points, so that in general the mass of the particle will not be the same for all points on its path. Plainly then, we cannot expect to arrive at a conformally invariant-system of dynamics so long as the mass of a particle is considered a fixed quantity belonging autonomously to the particle.

We have been concerned[6,7] since 1964 in developing a conformally invariant theory of dynamics and of gravitation. The central idea of this work is that the mass of a particle m say, is given by

$$m(X) = \lambda M(X) \tag{2}$$

where λ is a coupling constant belonging to the particle itself and $M(X)$ is a "mass field" generated by all the particles in the universe. In a very general way we can write

$$M(X) = \sum_b \int P(X, B)\, db \tag{3}$$

the summation being over all particles b, the line integral being over the world line of b, and $P(X, B)$ being a propagation function from point B at db to the field point X. To avoid self-action problems in the classical theory it is convenient to add the refinement that in determining the mass of any particular particle a this particle is omitted in the summation with respect to b. A comparable situation occurs in classical electrodynamics.

Mathematically we require that when all lengths on the right hand side of (3) are changed by the function Ω the resulting mass field is multiplied by $\Omega^{-1}(X)$. This requirement is met by putting

$$P(X, B) = \lambda \tilde{G}(X, B) \tag{4}$$

where $\tilde{G}(X, B)$ is the time-symmetric elementary solution of

$$\Box_x \tilde{G}(X, B) + 1/6\, R(X)\tilde{G}(X, B) = [-g(B)]^{-\frac{1}{2}}\, \delta_4(X, B) \tag{5}$$

in which $\Box_x$ is the d'Alembertian operator, R the scalar Riemannian curvature and δ_4 the 4-dimensional Dirac delta-function. So far as we are aware the wave-equation (5) was first considered with respect to its conformal properties by Penrose.

The action of particle a is now

$$-\lambda^2 \int da \,.\, \sum_{b \neq a} \int \tilde{G}(A, B)\, db \tag{6}$$

and the total action of all particles is given by

$$S = -\lambda^2 \sum\sum_{a \neq b} \iint \tilde{G}(A, B)\, da\, db \tag{7}$$

Subtleties arise for a universe with infinitely many particles concerning the finiteness of (7), but the situation in this respect is not really different from the usual theory.

Enough has already been done to make the Dirac equation conformally invariant, including the mass term. To obtain conformally invariant gravitational equations we work with the usual condition that S be stationary for small changes of the metric tensor, $\delta S = 0$ for $g_{ik} \rightarrow g_{ik} + \delta g_{ik}$. The resulting equations are different and more complicated than those of Einstein. For cosmology however, and for all problems concerned with weak gravitational fields, the equations can be reduced to the Einstein form with an exceedingly high degree of approximation. The situation in regard to the classical tests of general relativity—the bending of light, the perihelion of Mercury—is therefore exactly the same as with Einstein. To effect the reduction to Einstein's equations one chooses the conformal frame so that $M(X)$ becomes independent of X. If, to begin with, $M(X)$ varies with X we simply choose $\Omega \propto M$. Particle masses are independent of X in this particular conformal frame, just as we normally take them to be.

The gravitational 'constant' comes out to be given by

$$G = \frac{3}{4\pi M^2} \tag{8}$$

Again in the special conformal frame in which M is constant, G is indeed a constant. It is interesting that G is necessarily positive, so that weak gravitational fields are necessarily attractive.

It should be added that the gravitational equations do not reduce to the Einstein form in certain cases where the field is strong. The discussion of "black holes" is greatly changed, for example.

Because the gravitational theory is conformally invariant we can now effect a major simplification in discussing the isotropic homogeneous models usually employed in cosmology. These models are usually discussed with respect to the Robertson–Walker form of the line element

$$ds^2 = d\tau^2 - Q^2(\tau)\left[\frac{d\rho^2}{1 - k\rho^2} + \rho^2(d\theta^2 + \sin^2\theta\, d\varphi^2)\right] \tag{9}$$

Here τ is cosmic time and $Q(\tau)$ is the expansion function. The coordinates ρ, θ, φ are spherical polars for an observer at the origin. The constant k can be zero or ± 1. $k = 0$ gives the Einstein de Sitter model, $k = +1$ is closed, $k = -1$ hyperbolic. It is well known that a conformal transformation function Ω can be found that changes the line element (9) into flat Minkowski space.

It is not usually possible to take advantage of this geometrical simplification because the physical theory is not conformally invariant. Here, however, the physics is conformally invariant and we can transform to Minkowski space. When we have done so the theory is as good as it was in the Robertson–Walker representation. It turns out that while the $k = 0$ case is still homogeneous in the Minkowski space

$$dt^2 - dr^2 - r^2(d\theta^2 + \sin^2\theta\, d\varphi^2) \tag{10}$$

the cases $k = \pm 1$ are simply localized clouds of no special interest.

The full Minkowski space, $t < 0$ as well as $t > 0$, can be used. There is, however, nothing corresponding to both halves of Minkowski space in the Robertson–Walker representation. The reason for this emerges when we make the inverse transformation from (10) to (9). The inverse transformation has a singularity at $t = 0$, and this introduces a singularity at $\tau = 0$ in (9). This singularity, which now emerges as a mathematical construct, is usually interpreted as the 'origin' of the universe. There need be no 'origin' in the Minkowski representation.

Taking the particle masses to be constant in the Robertson–Walker frame, and applying Ω^{-1} to the mass, Ω being chosen for transformation from (9) with $k = 0$ to (10), we obtain

$$m = (\text{constant}).\, t^2 \tag{11}$$

The particle masses are a quadratic function of cosmic time t in the Minkowski representation. Even though space is now flat this leads to a redshift effect. Remembering that atomic sources emit frequencies proportional to m, we see that such a source at distance r from an observer at $r=0$ would be judged to have a redshift given by

$$1 + z = \left(\frac{t}{t - r}\right)^2 \tag{12}$$

the time of observation being t and the time of emission $t-r$.

A standard galaxy may be considered to have a specified number of stars each containing a specified number of particles arranged in the same way with respect to chemical composition. The luminosity of a star has dimensionality L^{-2}. This is proportional to m^2, and therefore by (11) proportional to the fourth power of the cosmic time. The bolometric flux from such a standard galaxy observed at time t is therefore

$$\frac{L(t-r)}{4\pi r^2} \propto \frac{(t-r)^4}{r^2} \tag{13}$$

where $L(t-r)$ is the luminosity of the galaxy at the time $t-r$ of emission of the observed radiation. Eliminating r from (13) with the aid of (12) we find that the bolometric flux depends on z according to

$$\frac{1}{1+z} \cdot \frac{1}{[\sqrt{1+z} - 1]^2} \tag{14}$$

which is the classic Hubble redshift-magnitude relation for the Einstein de Sitter model. The reader will note how much more easily we have arrived at (14) than in the usual treatment.

The present treatment is still deficient in one important respect. The form of equation (11) for the particle mass was obtained for the Minkowski representation by assuming constant mass in the Robertson–Walker representation. To make the argument complete it is essential to show that (11) follows from an explicit evaluation of (2), (3), (4), (5). This evaluation turns out to involve interesting problems which will not be considered here, since details have been given elsewhere[8]. It is convenient to choose the length unit L so that the particle density is everywhere unity. We then obtain from (2), (3),(4), (5),

$$m = \tfrac{1}{2}\lambda^2 t^2 L^{-3} \tag{15}$$

which has the required quadratic dependence on t.

Expanding (12) for $r/t \ll 1$ we obtain

$$z \simeq 2r/t \tag{16}$$

Since the Hubble constant H is defined by $z=Hr$ for small r, we have

$$t = 2H^{-1} \tag{17}$$

for the present epoch. Empirically $H^{-1} \simeq 2 \times 10^{28}$ cm. Empirically also, if we consider the particles to be hydrogen atoms, the average particle density n is given by

$$n = L^{-3} \simeq 2.5 \times 10^{-6}\ \text{cm}^{-3} \tag{18}$$

From (17), (18), and $H^{-1} = 2 \times 10^{28}$ cm,

$$nt^3 \simeq 10^{80} \tag{19}$$

This cosmological number is dimensionless, and is the number of particles contributing to the mass field $M(X)$ at the present epoch.

The dimensionless number λ^2 has no effect on gravitation, since it multiplies the action (7), and therefore disappears entirely from $\delta S=0$, $g_{ik} \to g_{ik} + \delta g_{ik}$. But the value of λ^2 has a profound effect on the behaviour of individual particles when electromagnetic terms are added to the action. The value of λ^2 can be obtained empirically in the following way. Applying (15) to protons,

$$m_p = \tfrac{1}{2}\lambda^2 t^2 L^{-3} = 2\lambda^2 H^{-2} L^{-3} \tag{20}$$

for the present epoch. The empirical values (18) for L^{-3} and $H^{-1} \simeq 2 \times 10^{28}$ cm then lead to

$$m_p \simeq 2 \times 10^{51}\ \lambda^2\ \text{cm}^{-1} \tag{21}$$

The laboratory value of the Compton wavelength of the proton m_p^{-1} is $\sim 10^{-14}$ cm, so that empirically $m_p \simeq 10^{14}$ cm^{-1}, and

$$\lambda^2 \simeq 5 \times 10^{-38} \tag{22}$$

is therefore required by atomicity.

The remarkable fact now emerges that

$$\lambda^2 (nt^3)^{\frac{1}{2}} = O(1) \tag{23}$$

On the left λ^2 apparently belongs autonomously to the particle, whereas $(nt^3)^{\frac{1}{2}}$ is a cosmological number. Apparently also λ^2 and n are independent of t, so that (23) would seem only a coincidence of the present epoch. This is the usual view.

At this point we reach our watershed. How if we require that (23) shall apply at all t? Then λ^2 or n, or both, must depend on t. This was the proposal of Dirac[5]. It is clear that (23) can be maintained by $n\lambda^4 \propto t^{-3}$, which allows many possible combinations of n, λ^2. Here we discuss a possibility which involves no change in the mass field $M(X)$. This possibility is similar in some respects to the case discussed by Dirac but is different in other respects. According to (2), (3), (4),

$$M(X) = \lambda \sum_b \int \tilde{G}(X, B)\, db \tag{24}$$

from which it is not hard to see that provided the product λn is independent of t there will be no change in $M(X)$. From

$$\lambda \propto n^{-1},\ n\lambda^4 \propto t^{-3} \tag{25}$$

we then get

$$\lambda \propto t^{-1},\ n \propto t \tag{26}$$

Moreover, the total action (7) is the same as before. It is as if more particles are being created by arranging that the masses of individual particles, $m=\lambda M$, shall increase linearly with t, instead of as t^2. In so far as we discuss the gravitational aspects of cosmology in terms of a smooth fluid, instead of in terms of discrete particles, nothing is changed. The interaction of each particle with the mass field M weakens as t^{-1}, but there are more particles because of $n \propto t$, and the total interaction of all particles remains the same.

An important by-product of these considerations is that the universe now becomes a perfect absorber in the sense of Wheeler and Feynman. The original Einstein de Sitter model did not meet the perfect absorber condition, which in our view is a fatal flaw of this model. The present model is not the steady state model, but by moving some way towards the steady state model, by permitting $n \propto t$, we have managed to meet an essential electrodynamic requirement.

It is possible to make a conformal transformation $\Omega \propto t$ from the Minkowski representation to a representation in which individual particle masses are constant, as we consider them to be in all terrestrial and astrophysical problems. The line

$$ds^2 = dT^2 - 2T[dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\varphi^2)] \tag{27}$$

which is the same as the line element of a radiation-dominated Friedmann model. The present model, however, is not radiation-dominated and it has a gravitational 'constant' that varies with time. In this representation the mass field M is proportional to $T^{\frac{1}{2}}$ and G, given by (8), is proportional to T^{-1}. Such a variation would have a profound effect on astrophysics and geophysics.

Stars would be much brighter at early T than we usually suppose them to be. There are advantages to be gained from such behaviour, but $G \propto T^{-1}$ may be too drastic[9], for example, in its effect on the past luminosity of the Sun. On the other hand, there are aspects of geophysics that seem as if they would be greatly helped by this kind of dependence. Steadily weakening gravity would gradually release the interior of the Earth from compression. It can be calculated[8] that the radius of the Earth would be required to increase at about 10 km per 10^8 yr. There is no possibility of this expansion being resisted by the crust, which must be cracked open repeatedly to make way for new surface material. At all times there would be an excess upward force on the crust at the limit of its strength. The possibility of large horizontal pressure differences, of order 10^9 dyne cm^{-2}, also exists, provided in particular regions that excess pressure is conveyed to the immediate subsurface by fluid material.

We are reminded in this connexion of the old controversy concerning continental drift. Our impression is that, while modern evidence shows unequivocally that drift actually takes place, the early calculations demonstrating the need for exceedingly large forces, really remain valid. If this is so, we would be inclined to think that some such behaviour of G as is given by our model becomes essential for an understanding of the geophysical evidence.

It may be added that Shapiro, Smith, Ash, Ingalls and Pettengill[10] have recently placed an observational upper limit of 4×10^{-10} yr^{-1} on $\dot{G}/G$. The variation expected here, with a Hubble constant of 5×10^{-11} yr^{-1}, would be 10^{-10} yr^{-1}.

We turn finally to the problem of how we should interpret $\lambda \propto t^{-1}$ in relation to the redshift problems discussed at the outset. The proportionalities $n \propto t$, $\lambda \propto t^{-1}$ maintain cosmological homogeneity. We must contemplate, however, that these proportionalities represent the smoothed-out effect of changes that could possess spatial irregularities. There could be spatially adjacent variations of λ. This we feel to be the kind of concept needed to come to grips with the problem of redshift anomalies. The smoothed-out homogeneous behaviour of the particle masses gives the normal cosmological redshift, obeying the usual Hubble relation. Local variations give anomalies. A more rapid decrease locally of λ is required in order that the anomalies be in the sense of an increased redshift. One can speculate that such variations occur as a consequence of the physical conditions in regions of strong gravitational fields. These are characterized by the condition that local particles make a contribution to the total mass field M that is comparable with the contribution of particles at cosmological distances. Such localities may be able to produce their own environmental conditions, leading to the local variations of λ.

Received July 27, 1971.

1 Arp, H. C., *Astrophys. Lett.*, **7**, 221 (1971).
2 Zwicky, F., *Naturwissenshaften*, **29**, 344 (1956).
3 Sargent, W. L. W., *Astrophys. J. Lett.*, **153**, L135 (1968).
4 Gunn, J., *Astrophys. J. Lett.*, **164**, L113 (1971).
5 Dirac, P. A. M., *Nature*, **139**, 323 (1937).
6 Hoyle, F., and Narlikar, J. V., *Proc. Roy. Soc.*, A, **294**, 138 (1966).
7 Hoyle, F., and Narlikar, J. V., *Ann. Phys.*, **62**, 44 (1971).
8 Hoyle, F., and Narlikar, J. V., *Mon. Not. Roy. Astron. Soc.* (in the press).
9 Pochoda, P., and Schwarzschild, M., *Astr. J.*, **139**, 587 (1964).
10 Shapiro, I. I., Smith, W. B., Ash, M. B., Ingalls, R. P., and Pettengill, G. H., *Phys. Rev. Lett.*, **26**, 27 (1971).

Origin of the Alps and Western Mediterranean

K. J. HSÜ
Geological Institute, ETH, Zürich

The origin of the Alpine geosyncline may have been related to the eastward motion of Africa relative to Europe during the Mesozoic, and orogenesis in western Europe could have been caused by a westward and then a northward movement of Africa.

In contrast to mountains such as the Himalayas and the Andes, the Alpine chain snakes its way through Europe with numerous bends, convergences and divergences[1,2]. Compared with the Atlantic, the opposite shores of which fit together almost perfectly, the Mediterranean is bounded by three peninsulas on the north, which run at right angles to the trend of the sea. To make matters worse the Mediterranean Ridge has proved to be a compressional feature[3], quite distinct from the normally extensional mid-ocean ridges.

The complex state of the Mediterranean reflects its past history and conventional interpretations have related the origin of the Alps and the Mediterranean to the interaction between the European and African plates. In many of these interpretations a broad, east–west sea (the Tethys Sea) was placed between the two continents (Fig. 1), and the Alpine folding was postulated as a consequence of the northward march of Africa[4,5]. Such a broad new picture left many of the intricacies unexplained and the important geological fact that a deep oceanic seaway did not exist in western Europe (the Alps, Apennines, and Betic Cordilleras) before the Jurassic was ignored.

Geometrical Analysis

Staub[4] considered the winding Alpine trend as a trait inherited from the European Hercynian system and thereby evaded the issue. Argand[6], however, presented a brilliant geometrical analysis and invoked continental drift and rotation of microcontinents to account for the puzzling geography of the western Mediterranean. But his kinematical analyses were based on geological data alone and have been proved untenable by recent geophysical and oceanographical discoveries. This article is an attempt to modify Argand's scheme; if the drifting chronology is revised, the geometrical displacements postulated by Argand acquire a different geological significance.

Argand[6] postulated that Africa was pushed over Eurasia in the early Tertiary, and formed the Alpine and Carpathian mountains while Eurasia was still attached to North America. After the climax of the Alpine deformation in the Oligocene,