PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A
MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

The Cosmological Constant [and Discussion]

S. W. Hawking

Phil. Trans. R. Soc. Lond. A 1983 **310**, 303-310
doi: 10.1098/rsta.1983.0092

Phil. Trans. R. Soc. Lond. A **310**, 303–310 (1983)
Printed in Great Britain

The cosmological constant

BY S. W. HAWKING, F.R.S.

University of Cambridge, Department of Applied Mathematics and Theoretical Physics, Silver Street, Cambridge CB3 9EW, U.K.

The cosmological constant is the quantity in physics that is most accurately measured to be zero: observations of departures from the Hubble law by distant galaxies place an upper limit of the order of 10^{-120} in dimensionless units. On the other hand, the various symmetry breaking mechanisms that we believe are operating in the Universe would give an effective cosmological constant many orders of magnitude larger, unless they are incredibly finely balanced. One answer would be to appeal to the anthropic principle, but a more attractive possibility is that there is a phase transition in $N = 8$ supergravity to a foam-like state which breaks supersymmetry and which appears flat on scales larger than the Planck length.

The history of the cosmological constant goes back to the seventeenth century. Newton realized that the so-called 'fixed stars' were like the Sun and he postulated that they were distributed approximately uniformly throughout space. This raised a problem: how could they remain at roughly constant distances from each other if they were attracting each other according to his law of gravity? Why did they not all fall together? One possible solution was to add a repulsive (in both senses of the word) 'cosmological' term to the Newtonian equation:

$$\nabla^2\phi + \Lambda = 4\pi G\rho. \qquad (1)$$

This would not make a significant difference to orbits in the solar system but it would allow static cosmological solutions with a uniform, non-zero average density. These solutions were unstable, however, because, if the density were slightly higher than average in some region, the gravitational attraction would dominate over the repulsion produced by the Λ term and the region would contract. Similarly, if the density were slightly less than average, the repulsion would win and would cause the region to expand indefinitely.

A static universe was still the accepted cosmological model when Einstein proposed the general theory of relativity in 1915. He therefore added a similar cosmological term to his field equation:

$$R_{ab} + \tfrac{1}{2}g_{ab}R + \Lambda g_{ab} = 8\pi G T_{ab}, \qquad (2)$$

(I shall use units in which $c = 1$). This again allowed static cosmological solutions with a uniform non-zero density although, as before, they were unstable. However, a few years later, it was discovered that other galaxies were receding from our own with a velocity that increased with their distance from us. The static model of the Universe therefore had to be abandoned in favour of an expanding one. This removed the original reason for introducing the cosmological term. If one now regarded it as an adjustable parameter in the theory, one could attempt to measure it by observing departures from linearity in the magnitude redshift plot for distant galaxies (Sandage & Tammann 1982), figure 1.

If one assumes that the Universe is spatially homogeneous and isotropic (which seems to be a good approximation on large scales), then it is described by a Robertson–Walker metric with scale factor R. The Einstein equations give

$$3H^2 = 3\dot{R}^2/R^2 = 8\pi G\mu + \Lambda - k/R^2, \tag{3}$$

$$q_0 H^2 = \ddot{R}/R = -\tfrac{4}{3}\pi G(\mu + 3p) + \Lambda, \tag{4}$$

where μ is the average energy density and p is the average pressure. One can determine the Hubble constant H from the magnitude-redshift plot if one knows the intrinsic luminosity of a typical galaxy. One finds a value for H between 10^{-10}/year and 5×10^{-11}/year. If one assumed that all the galaxies had the same intrinsic luminosity, one could determine q_0 by measuring departures from linearity in the magnitude-redshift plot. One would expect however that the luminosities of the galaxies would have changed with time but one does not know how much. Nevertheless, one can be confident that $-5 < q_0 < 5$. One can also place a lower bound of 3×10^{27} cm on the scale factor R from the fact that one sees distant galaxies which are apparently randomly distributed. If one now uses the fact that the average pressure p in the universe is small, one can place an upper bound on $|\Lambda|$ of $10^{-54}\,\text{cm}^{-2}$.

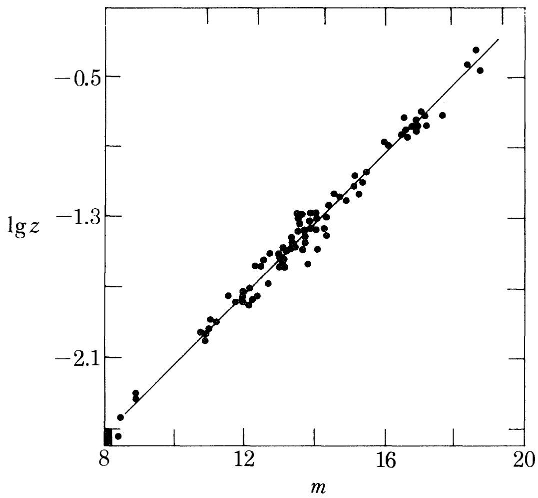

FIGURE 1. The magnitude redshift plot for the brighest galaxies in clusters of galaxies (see Sandage & Tamman 1982).

From the point of view of the classical general relativity, Λ is an *ad hoc* addition to the theory and these is no natural length scale with which to compare it. Thus, there is no reason to give it any particular value other than zero. However, when one takes quantum mechanics into account, the gravitational constant G has dimensions of mass^2 or length^{-2}. In other words, one can use units in which $\hbar = 1$ and one can write $G = M_P^{-2} = l_P^2$, where M_P is the Planck mass *ca.* 10^{19} GeV and l_P is the Planck length *ca.* 10^{-33} cm. One can then express the observational upper bound in the dimensionless form

$$\Lambda/M_P^2 < 10^{-120}. \tag{5}$$

This makes Λ the quantity in physics that is most accurately measured to be zero. By contrast, the observational upper limit on the mass of the photon obtained from spacecraft measurement of the Earth's magnetic field is only

$$m_\gamma^2/m_e^2 < 10^{-48}. \tag{6}$$

Yet we do not regard the mass of the photon as an adjustable parameter which just happens to have a very low value. Rather, we invoke a physical principle, gauge invariance, to make it exactly zero. The trouble is that there does not seem to be any similar principle or symmetry to which one can appeal to make Λ zero. In fact there seem to be several effects which would be expected to give rise to an effective value of Λ much larger than the observational upper limit (5) unless they cancel each other out to a very high degree of accuracy.

In a quantum state, such as the vacuum state, which is approximately invariant under local Lorentz transformations, the expectation value of the energy momentum tensor must be proportional to the metric. This gives rise to an effective cosmological constant Λ:

$$\Lambda = 8\pi M_{\mathrm{P}}^{-2} \langle T_{00} \rangle. \tag{7}$$

In other words, the vacuum energy produces an effective cosmological constant. There are a number of effects that will contribute to the vacuum energy. To start with, the zero point fluctuations in each mode will contribute $\pm\frac{1}{2}\omega$ where the plus sign is for bosons and the minus sign is for fermions. Unless the number of boson and fermion degrees of freedom were equal (a possibility which I will come to), one would have to cut off the fluctuations at some frequency ω_0. One would then have an effective Λ of the order of $\omega_0^4 M_{\mathrm{P}}^{-2}$. The most natural value of the cut off would be $\omega_0 = M_{\mathrm{P}}$. This would violate the upper bound (5) by 120 orders of magnitude.

Even if the vacuum fluctuations were renormalized to zero, one would still get large changes in the vacuum energy when symmetries were broken. The contribution to Λ would be of the order of $\mu^4 M_{\mathrm{P}}^{-2}$ where μ is the energy at which the symmetry is broken. We have good experimental evidence that chiral symmetry is broken around 200 MeV and that electro-weak symmetry is broken around 100 GeV. In addition there are indications that there may be a grand unified symmetry between strong and electroweak interactions which is broken somewhere between 10^{15} and 10^{19} GeV. It is very difficult to believe that the zero in the renormalization of the vacuum energy could have been chosen so exactly as to cancel out the contributions of all the symmetry breakings to better than one part in 10^{40}.

There is another possible symmetry apart from those mentioned above. It is supersymmetry, which relates fields of different spin. In a supersymmetric theory the number of boson and fermion degrees of freedom are equal so the infinite contributions to Λ cancel out. However, supersymmetry would also imply that all particles have the same mass, and clearly they don't. Thus, if the fundamental theory is supersymmetric, the supersymmetry must be broken at some energy greater than about 1000 GeV. This would give a positive contribution to Λ which would exceed the observational limit by 56 orders of magnitude. It might be possible to balance this with a negative contribution from some other form of symmetry breaking but this would require exceptionally fine tuning.

One possible explanation of the smallness of the cosmological constant would be to invoke the anthropic principle. One could imagine that there were very many different universes with different values of the cosmological constant. Only in those in which the cosmological constant was very small would it be possible for intelligent life to develop and ask the question: why is the cosmological constant so small? This idea will be discussed further in the lecture by Brandon Carter. Here, I will describe a different possibility, namely, that there is a phase transition to a state in which space–time is 'foam-like' on scales of the Planck length but appears smooth and nearly flat with zero cosmological constant on larger scales.

I shall assume that the fundamental theory of the Universe is $N = 8$ gauged extended supergravity (de Wit & Nicolai 1981). The reasons for this are that, first, supergravity theories seem to offer the only hope of quantizing gravity without non-renormalizable divergences. It is known that the $N = 8$ theory is finite at the one and two loop level (apart from a possible topological divergence that I shall describe below). There is a possible counter term with the right symmetry at the three loop level but it is not known whether it actually appears with a non-zero coefficient. If it does, it seems that one will have to go to a higher derivative theory with all the problems that has with ghost states: lack of unitarity and runaway solutions. The other attraction of supergravity theories is that they seem to provide the only way of unifying gravity with the other interactions of physics such as the strong and weak forces and electromagnetism. The $N = 8$ gauged theory is the largest of all such supergravity theories and the most likely to be finite at all numbers of loops. On the other hand, there seem to be several obstacles to accepting $N = 8$ gauged supergravity as the ultimate theory of the observed Universe.

(i) The ground state of the $N = 8$ theory is anti-de Sitter space which has a large negative cosmological constant Λ of order $-e^2M_P^2$ where e is the coupling constant of the SO(8) gauge fields.

(ii) The theory is supersymmetric whereas the observed Universe is not.

(iii) Although the $N = 8$ theory contains a large number of different particles, it does not seem to contain enough. In particular, the SO(8) gauge group does not contain the SU(3) $\times$ SU(2) $\times$ U(1) group which seems to be observed.

I shall show how a phase transition to a foam-like state could answer objections (i) and (ii). It might also answer (iii), though that remains to be determined.

Gauged extended supergravity theories for $N \geqslant 4$ contain scalar fields with a potential that has extremum points at negative values of the potential. One of these extrema gives rise to a supersymmetric solution which is anti-de Sitter space with constant values of the scalar fields. One might think that this solution would be unstable because the extremum point is a local maximum and the potential is unbounded below. Nevertheless, one can show that the solution is stable both to small fluctuations (Breitenlohner & Freedman 1982) and, indeed, to all fluctuations (Gibbons *et al.* 1982) provided that the fields obey certain boundary conditions (Hawking 1983). There are other extrema of the scalar potential (Warner 1983*a*, *b*) that give rise to anti-de Sitter solutions with less supersymmetry or with no supersymmetry at all. However, the result of Gibbons *et al.* (1982) indicates that there cannot be any quantum tunnelling from the fully supersymmetric solution to these other solutions. One can understand this in the following way: the tunnelling would proceed by a 'bubble solution' which consisted of a region of the new solution surrounded by the fully supersymmetric solution. There would be an energy gain which would be proportional to the volume of the region and an energy loss which would be proportional to the area of the boundary of the region. In flat space one can make the ratio of volume to area indefinitely large by making the region large enough. Thus, it is always possible to tunnel. However, in anti-de Sitter space, the ratio of volume to area is bounded. Thus, it may not be possible to tunnel if certain inequalities are satisfied (Coleman & de Lucia 1980) and these inequalities hold in supersymmetric theories (Weinberg 1982).

The arguments given above indicate that it is not possible to break supersymmetry by tunnelling to one of the other extrema of the scalar potential. In $N \leqslant 4$ supergravity one can break supersymmetry by adding supermatter fields. One can even arrange it so that a positive contribution from the matter fields exactly cancels the negative cosmological constant of the gauged supergravity theory though this requires very fine tuning. However, there is no supermatter for

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

$N > 4$ so one cannot break $N = 8$ supersymmetry in this way. If one added $N = 4$ supermatter to $N = 8$ supergravity, one would obtain a theory that had at most four supersymmetries and that would be likely to have divergencies even at one loop. Thus the only possible way to break $N = 8$ supersymmetry without introducing unrenormalizable divergencies would seem to be to appeal to radiative corrections. However one would expect that there would be non-renormalization theorems which would prevent the breaking of supersymmetry by radiative corrections unless they were infinite. But, as was said before, the theory would be non-renormalizable if the radiative corrections were infinite. The one exception to this statement is a possible one-loop divergence that is proportional to a topological invariant.

The one-loop divergences of gauged supergravity theories have been analysed by Christensen *et al.* 1980 in the Euclidean régime, i.e. for positive definite metric. There are two possible terms:

$$A\chi + B\int \Lambda^2\,\mathrm{d}v, \tag{8}$$

where χ is the Euler number of the space. The coefficient B is zero for $N \geqslant 5$ (cf. Allen & Davis 1983). This is related to the nonexistence of supermatter for $N > 4$. The coefficient A is $(3-N)$ for $N \geqslant 3$. There is an alternative formulation of the ungauged $N = 8$ theory which is obtained by dimensional reduction from the $N = 1$ theory in eleven dimensions (Cremmer & Julia 1979). In this, 7 of the 70 scalar fields are replaced by antisymmetric tensor fields that have the same number of physical degrees of freedom. The coefficient A for this theory is zero (Siegel 1981; Duff 1981). However, this theory cannot be invariant under the group SO(8) because 7 is not a representation of SO(8). In fact it seems that it is a version of the ungauged $N = 7$ theory which has the same physical degrees of freedom as the $N = 8$ theory. It seems that in any version of the $N = 8$ theory which has SO(8) as a gauge group, the coefficient A will equal minus five, and certainly this is true of the only gauged $N = 8$ theory that we have, the one given by de Wit & Nicolai (1981).

The fact that A is non-zero means that one will have to add a term $-k\chi$ to the effective action where k is a scale dependent topological coupling constant. The fact that A is negative implies that topological fluctuations will be more important on small scales (the opposite of asymptotic freedom):

$$k = k_0 - 5\ln(l/l_{\mathrm{P}}), \tag{9}$$

where l is a typical length scale of the topological fluctuation. At first sight, it would seem that the metric would have more and more topological fluctuations on smaller and smaller length scales and that space–time would not be a smooth manifold but would be something like a fractal. We do not know how to formulate such a theory and, even if we did, there would be no reason why space–time should have four dimensions rather than three and a half or π. Perhaps fortunately, it seems that there is a self-limiting effect which suppresses fluctuations on scales less than the Planck length. I shall give a brief outline of the idea here and more detail in another publication.

The Euclidean action for the gravitational field is not bounded below because the kinetic term for the conformal degree of freedom is negative definite (Gibbons *et al.* 1978). In order to deal with this, one has to divide the metrics in the path integral into conformal equivalence classes and perform the path integral over the conformal factor Ω along a contour in the complex Ω plane which crosses the real axis (cf. Hartle & Hawking 1983). One can deform the Ω contour so that it crosses the real axis at the lowest saddle point, i.e. the maximum value of the Euclidean action on the real Ω axis (there will always be a maximum if the cosmological constant is negative). One would expect that the dominant contribution to the path integral over the conformal factor would come from the saddle point.

Consider a metric which has a roughly uniform distribution of topological fluctuations. Under a constant conformal transformation Ω the action of the non-gravitational fields will be unchanged and the Euclidean action I of the gravitational field of a region of this metric will be:

$$I = C_1\Omega^2 - C_2\Omega^4 + C_3 \ln(\Omega/\Omega_0). \tag{10}$$

The first term is the gravitational action $-(16\pi)^{-1} M_P^2 \int R\, dv$. The constant C_1 will be roughly proportional to the square root of the Euler number χ of the region (Hawking 1978). The second term is the contribution of the cosmological constant. The constant C_2 will therefore be proportional to e^2. The third term is the contribution of the topological term in the action. The constant C_3 will be proportional to the Euler number χ and Ω_0 is related to the renormalization constant k_0.

If one neglects the topological term, the action will have a positive maximum value on the real Ω axis proportional to χe^{-2}. Thus, topological fluctuations will have higher actions and so will be suppressed. However, if one includes the topological term, the maximum value will occur for

$$\Omega^2 \approx C_4 e^{-2} \chi^{\frac{1}{2}} (1 + C_5 e), \tag{11}$$

where C_4 and C_5 are constant. If e is small, the maximum value of I will still be positive and proportional to χe^{-2}. Thus, topological fluctuations would still be suppressed. However, if e were greater than some critical value e_0, the maximum value of I would be negative and proportional to χ times some function of e. In this case, it would be more favourable to have topological fluctuations than not to have them and they would occur predominantly on the scale given by (11), i.e. about one topological fluctuation or 'bubble' per Planck volume. It therefore seems that the $N = 8$ theory may have two different régimes or phases: if $e < e_0$, the ground state is anti-de Sitter space and supersymmetry is unbroken; but if $e > e_0$, the ground state will contain a large number of topological fluctuations, rather like the confinement phase of Yang–Mills theory. There is, however, a difference from Yang–Mills theory: the coupling constant e is not renormalized and is scale independent. One would therefore expect the $N = 8$ theory to be either in the anti-de Sitter phase or in the space–time foam phase on all length scales according to whether e is less or greater than the critical value.

What would the space–time foam phase look like? To start with, it would almost certainly break supersymmetry because it is only a very few special spaces, like anti-de Sitter space, which possess Killing spinors, i.e. supercovariantly constant spinor fields. However, it would seem that it would break supersymmetry without giving a mass to the spin $\frac{3}{2}$ fields as would normally happen. Equation (11) indicates that the topological fluctuations or bubbles would occur predominantly on the Planck length scale. Thus, one might expect that on larger scales, space–time would appear smooth. However, the presence of the bubbles would mean that the apparent curvature on large scales could be very different from that on small scales. One can see this effect by considering solutions of the Einstein equations with a negative cosmological constant. A theorem of Yau (1977) implies that there exist large numbers of such solutions with arbitrarily high Euler number. In these, the volume integral of the square of the Weyl tensor will be of the order of χ. The Weyl tensor will be randomly orientated on the scale of the bubble length and will introduce shear, and hence convergence, on congruencies of geodesics. On a scale large compared with the bubbles it will therefore appear to have the same effect as a positive contricution to the Ricci tensor of order $\rho^{\frac{1}{2}}$ where ρ is the Euler number density per unit volume. Thus, the bubbles will tend to cancel out the negative Ricci tensor, which arises from the potential of the scalar fields. But why should they cancel it out to one part in 10^{120}? The answer is that, if it is favourable to have bubbles, then it is more favourable to have more bubbles. Thus, the density of bubbles tries to be as high as it can. However, if the apparent Ricci tensor on large scales becomes positive, then the whole space curls up on itself and has a small euclidean four-volume. In this case, the total number of bubbles would be limited. The most favoured case therefore is when the density of bubbles is such that the apparent large scale Ricci tensor is zero, i.e the apparent cosmological constant is zero.

The suggestion is therefore that the $N = 8$ gauged supergravity theory has a phase transition at a certain critical value of the gauged coupling constant e. Above that value, the ground state would be 'foam-life' with topological fluctuations on the Planck length scale but smooth and exactly flat on larger length scales. This bubble mechanism seems to be the only way of getting a zero effective cosmological constant without fine tuning. The foam would severely affect the propagation of elementary scalar particles but not that of low energy particles of spin half or higher (Hawking *et al.* 1980). The number of spin $\frac{1}{2}$ particles in the $N = 8$ theory is 56, which is large enough to accommodate all the observed particles. The observed spin 1 particles could not be the fundamental $N = 8$ particles because the group SO(8) does not contain the observed $SU(3) \times SU(2) \times U(1)$. Presumably, they would have to be composites of the fundamental spin 0 or $\frac{1}{2}$ particles. Such composite particles would probably not be much affected by the space–time foam.

References

Allen, B. & Davis, S. 1983 *Physics Lett.* B **124** B, 353.
Breiltenlohner, P. & Freedman, D. Z. 1982 *Ann. Phys.* **144**, 249.
Christensen, S. M., Duff, M. J., Gibbons, G. W. & Rocek, M. 1980 *Phys. Rev. Lett.* **45**, 161.
Coleman, S. & de Lucia, F. 1980 *Phys. Rev.* D **21**, 3305.
Cremmer, E. & Julia, B. 1979 *Nucl. Phys.* B **159**, 141.
de Wit, B. & Nicolai, H. 1981 *Physics Lett.* B **108**, 285.
Duff, M. J. 1981 Antisymmetric tensors and supergravity. In *Superspace and supergravity* (ed. S. W. Hawking & M. Rocek), pp. 381–402. Cambridge: University Press.
Gibbons, G. W., Hawking, S. W. & Petty, M. J. 1978 *Nucl. Phys.* B **138**, 141.
Gibbons, G. W., Hull, C. M. & Warner, N. P. 1982 The stability of gauged supergravity. (DAMTP Preprint).
Hartle, J. B. & Hawking, S. W. 1983 Wave function of the Universe. *Phys. Rev.* D (Submitted.)
Hawking, S. W., Page, D. N. & Pope, C. N. 1980 *Nucl. Phys.* B **170**, 283.
Hawking, S. W. 1978 *Nucl. Phys.* B **144**, 349.
Hawking, S. W. 1983 The boundary conditions in gauged supergravity. *Physics Lett.* B (Submitted.)
Sandage, A. R. & Tammann, G. A. 1982 H_0, q_0 and the local velocity field. In *Astrophysical cosmology* (ed. H. A. Bruck, G. V. Coyne & M. S. Longair). Pontificae Academiae Scientiarum Scripta Varia. **48**.
Siegel, W. 1981 The wraiths of graphs. In *Superspace and supergravity* (ed. S. W. Hawking & M. Rocek), pp. 151–164. Cambridge: University Press.
Warner, N. P. 1983*a* Some properties of the scalar potential in gauged supergravity theories. California Institute of Technology preprint.
Warner, N. P. 1983*b* Some new extrema of the scalar potential of gauged N = 8 supergravity. California Institute of Technology preprint.
Weinberg, S. 1982 *Phys. Rev. Lett.* **48**, 1776.
Yau, S. T. 1977 *Proc. natn. Acad. Sci. U.S.A.* **74**, 1798.

Discussion

H. B. Nielsen (*Bohr Institute, Copenhagen, Denmark*). Is it essential for one's understanding of the zeroness of the cosmological constant to have the $N = 8$ supergravity? The bubbles might occur in many gravity models. What are the essential assumptions?

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

S. W. Hawking. $N = 8$ supergravity is not essential. But it is necessary to the situation that I have described, in that the small scale value of the cosmological constant is negative because bubbles always add a positive contribution to the apparent large scale cosmological constant. In fact one can obtain a zero effective cosmological constant without appealing to topological fluctuations at all: one can replace the density of bubbles by a 3-index anti-symmetric tensor field. This has no dynamics: the field equations are that the curl is constant. This provides a constant effective contribution to the cosmological constant and a similar argument to that for bubbles shows that the probability is highest when the total cosmological constant is zero: if it is negative, the action is positive and the probability is exponentially small, if it is positive, the Euclidean volume of the space is limited and the action is bounded below. The most favourable case is when the total cosmological constant is exactly zero.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES