1/3 factor in the CMB Sachs-Wolfe effect

J. Hwang,[1,2] T. Padmanabhan,[3] O. Lahav,[2] and H. Noh[4,2]
[1]*Department of Astronomy and Atmospheric Sciences, Kyungpook National University, Taegu, Korea*
[2]*Institute of Astronomy, Madingley Road, Cambridge, United Kingdom*
[3]*Inter-University Centre for Astronomy and Astrophysics, Post Bag 4, Ganeshkind, Pune 411 007, India*
[4]*Korea Astronomy Observatory, San 36-1, Whaam-dong, Yusung-gu, Daejon, Korea*
(Received 17 July 2001; published 23 January 2002)

We point out that a pseudo Newtonian interpretation of the 1/3 factor in the Sachs-Wolfe effect, which relates the fluctuations in temperature and potential, $\delta T/T=(1/3)\delta\Phi$, is not supported by the general relativistic analysis. Dividing the full gravitational effect into separate parts depends on the choice of time slicing (gauge) and there exist infinitely many different choices. More importantly, interpreting the parts as being due to the gravitational redshift and the time dilation is not justified in the rigorous relativistic perturbation theory. We suggest regarding the 1/3 factor as a general relativistic result that applies in a restricted situation of adiabatic perturbation in the $K=0=\Lambda$ model with the last scattering occurring in the matter dominated era. For an isocurvature initial condition the corresponding result $\delta T/T=2\delta\Phi$ has a different numerical coefficient.

DOI: 10.1103/PhysRevD.65.043005 PACS number(s): 98.70.Vc, 98.80.Hw

I. INTRODUCTION

Some of the exact results in general relativity (GR) can be given intuitive interpretations in terms of Newtonian concepts even though the structures of these two theories are very different. (i) The simplest example is the ''derivation'' of the Schwarzschild radius by equating the Newtonian escape velocity to the velocity of light; one gets the correct numerical factor although the argument is not correct. (ii) The collapse time for a uniform density dust sphere under self-gravity computed by Newtonian theory turns out to be exactly equal to the general relativistic result. (iii) The gravitational force on a material particle located inside the empty region of a spherical shell of matter vanishes in Newtonian theory. This is usually explained by the fact that the force in Newtonian gravity falls as $1/r^2$ while the amount of matter intercepted by a cone with a fixed solid angle grows as r^2 thereby leading to the cancellation of forces due to opposite pairs of material in the shell [1]. Incredibly enough this result is true in GR as well although the force law in GR is not strictly $1/r^2$ while the material intercepted by a fixed solid angle does increase as r^2. (iv) At a more subtle level, one can obtain the time-time component of Einstein's equations for a FLRW (Friedmann-Lemaître-Robertson-Walker) universe from the Newtonian energy conservation argument if we take the potential energy of a spherically symmetric mass distribution to be $-(GM/R)$. The argument is again invalid in GR especially since Birkhoff's theorem is applicable only for empty regions outside a mass distribution. In all four cases mentioned above, these are accidental coincidences and the heuristic arguments have no deeper significance.

Recently, another pseudo Newtonian argument has surfaced to explain the origin of the 1/3 factor in the Sachs-Wolfe effect in cosmic microwave background (CMB) radiation [2]. The following heuristic argument is often seen in the literature including several textbooks [3–9]: $\delta T/T=(1/3)\delta\Phi=\delta\Phi-(2/3)\delta\Phi$ where the first term arises from the gravitational redshift whereas the second term comes from the time dilation of the temperature perturbation $\delta T/T=-\delta a/a=-(2/3)(\delta t/t)=-(2/3)\delta\Phi$ where a is the scale factor. For a similar argument see [10]. We point out in this paper that the above decomposition is motivated by a certain slicing condition (temporal gauge choice) out of infinitely many possible choices. More importantly, however, the interpretation of the terms as above cannot be explained even in that slicing or in any other slicing in the context of the correct relativistic analyses; see below Eq. (9). Hence, as the heuristic argument and interpretation are not supported by rigorous theory, we believe such are spurious and have at best the same status as the examples we mentioned in the first paragraph.

Before proceeding with a rigorous analysis of the situation, it is worth pointing out that there are some fundamental difficulties in providing *any* Newtonian interpretation of the Sachs-Wolfe effect. Given any metric in GR, the Newtonian limit can be rigorously established if the metric can be reduced to the form $g_{00}=-(1+2\Phi)$ and $g_{\alpha\beta}=\delta_{\alpha\beta}$ in some coordinate system. When one attempts to do this with the FLRW metric (see [11]), one finds that the procedure is valid only for a region of size much smaller than the Hubble radius $d_H(t)\equiv(\dot{a}/a)^{-1}$. This merely reiterates the fact that at scales bigger than the Hubble radius one requires the full machinery of GR and—in particular—one needs to grapple with issues of gauge. Since the Sachs-Wolfe effect arises from scales that are bigger than the Hubble radius at the epoch of recombination one can anticipate that *any* Newtonian interpretation will have problems. This conclusion is also supported by the fact that one can consistently choose a gauge in a perturbed FLRW universe in which $g_{00}=-1$. There is no Newtonian potential available in this gauge. The key reason for the above pseudo derivation to work is because one introduced an ill-defined quantity $\delta a/a$ and linked it with $\delta T/T$ at one end (by $Ta=\text{const}$) and with $\delta t/t$ at the other end (by $a\propto t^{2/3}$). There is no rigorous interpretation of this quantity $\delta a/a$ possible in Newtonian gravity, or for that matter in the gauge invariant treatments of FLRW perturbations. In handling the cosmological perturbation one can adopt a heuristic argument using δa. However, in this way, one often ends up with an equation that is wrong (especially in a medium with pressure).[1] This is not surprising because perturbing the background system necessarily loses some degrees of freedom compared with perturbing the full system.

II. TEMPERATURE ANISOTROPY IN AN ARBITRARY GAUGE

We begin by obtaining the expression for the temperature anisotropy without imposing any choice of gauge so that we can study the results in the most general setting. We consider a spatially homogeneous and isotropic metric with the scalar-type perturbations

$$ds^2 = -a^2(1+2\alpha)d\eta^2 - 2a^2\beta_{,\alpha}d\eta\, dx^\alpha + a_2[g^{(3)}_{\alpha\beta}(1+2\varphi) + 2\gamma_{,\alpha|\beta}]dx^\alpha dx^\beta. \quad (1)$$

This represents a fairly general perturbed metric and no specific gauge has been chosen. The variables α, β, φ, and γ are spacetime dependent scalar-type metric perturbations, and a vertical bar indicates a covariant derivative based on the background three-space comoving metric $g^{(3)}_{\alpha\beta}$. We introduce the variable $\chi \equiv a(\beta + a\dot{\gamma})$ which gives the shear of the normal hypersurface. φ is proportional to the perturbed curvature of the hypersurface, and α is the perturbed lapse function; we use the notation of [13,14]. The combination of χ and the rest of the variables used in the following is spatially gauge invariant [13]. The variables depend, however, on a temporal gauge (coordinate) transformation which corresponds to choosing the spatial hypersurface, i.e., the time slicing. Thus we have the freedom to impose a temporal gauge (slicing) condition which could be used as an advantage to handle the problems conveniently. The prime and the overdot indicate the time derivatives based on η and t, respectively, with $dt \equiv a\, d\eta$.

The most general expression of the Sachs-Wolfe effect from the scalar-type perturbation can be found in Eq. (15) of [14]. Ignoring the Doppler effects due to the observer's motion, and the emitting event (which is subdominant at large angular scales), the observable temperature anisotropy becomes

$$\left.\frac{\delta T}{T}\right|_O = \left.\left(\alpha_\chi + \frac{\delta T_\chi}{T}\right)\right|_E + \int_E^O (\alpha_\chi - \varphi_\chi)' dy, \quad (2)$$

where the integration is along the photon's null geodesic path from the emitted (E) epoch to the observed (O) epoch. The variables $\alpha_\chi \equiv \alpha - \dot{\chi}$, $\delta T_\chi \equiv \delta T + HT_\chi$, and $\varphi_\chi \equiv \varphi - H_\chi$ are gauge-invariant combinations.[2] φ_χ becomes φ in the zero-shear gauge (often called the Newtonian gauge) which sets $\chi \equiv 0$, etc. Using Bardeen's notation in [15] we have $\alpha_\chi \equiv \Phi_A$ and $\varphi_\chi \equiv \Phi_H$ [16]. The gauge invariance of a combination assures that there remains no gauge (coordinate) mode and that its value remains the same in any gauge. However, it does not guarantee that a certain gauge-invariant variable has an associated intrinsic physical meaning independent of the slicing condition. For example, there exist several variables φ_χ, $\varphi_{\delta T} \equiv \varphi + \delta T/T$, $\varphi_v \equiv \varphi - (aH/k)v$, where v is a velocity related variable and k is a wave number, etc., which are all gauge invariant. (An exception is φ_α based on the synchronous gauge which fixes $\alpha = 0$.) They all reduce to φ under the corresponding gauge condition which sets the variable in the subscript equal to zero. Although all these gauge-invariant variables (we can make infinitely many different combinations) are curvature variables in different time slicings (temporal gauges) we can safely regard them as completely different variables.

Equation (2) was derived from the geodesic equations in the spacetime of Eq. (1) *without* fixing the gauge condition and *without* using the gravitational field equations. In the literature, the two terms in the right-hand side (RHS) are often called the Sachs-Wolfe effect and the integrated Sachs-Wolfe effect, respectively. At this point it may be appropriate to quote a comment in [2] which looks prescient in the context of the main point of this paper: "We emphasize again that in a generic gravitational field one cannot distinguish gravitational redshifts from Doppler shifts by any standard recipe; thus our division of the equation...has only a heuristic significance."

We shall now assume that (i) the anisotropic stress can be ignored, so that we have[3] $\alpha_\chi = -\varphi_\chi$, (ii) $K = 0 = \Lambda$, and (iii) the medium is an ideal fluid with constant $w \equiv p/\mu$ (where p is the pressure and μ is the energy density) so that the growing solution of φ_χ remains constant in time.[4] Given all these assumptions the integrated Sachs-Wolfe term vanishes, so that

$$\left.\frac{\delta T}{T}\right|_O = (-\varphi + 2H\chi + \tfrac{1}{4}\delta_{(\gamma)})|_E, \quad (3)$$

where we used $\delta T/T|_E = (1/4)\delta_{(\gamma)}|_E$ with $\delta_{(\gamma)} \equiv \delta\mu_{(\gamma)}/\mu_{(\gamma)}$ denoting fractional energy-density perturbation of the photons. Now, the RHS is written without fixing the gauge yet; thus it is in a sort of gauge-ready form, but the sum is gauge invariant. In this form we can understand why and how such a decomposition into the intrinsic temperature perturbation and the gravitational redshift (or the time dilation) is dependent on the temporal slicing (gauge) condition of the spacetime. Although φ and χ have meaning as the perturbed curvature and shear of the normal three-hypersurface, and $\delta_{(\gamma)}$ looks like an energy-density perturbation of photons, these variables acquire such a meaning only after fixing the temporal gauge (time slicing) condition, where we have infinite choices. As mentioned before, the same variable evaluated in a different slicing (gauge) condition in general behaves as a completely different variable.

[1] This is in contrast with the perturbed Hubble parameter δH, which has rigorous geometric and kinematic meanings, and in fact is quite useful in handling the cosmological perturbation in a heuristic looking but fully rigorous manner [12].

[2] For the gauge transformation properties, see Eq. (2) in [14].

[3] See Eq. (8) in [14].

[4] See Eq. (18) in [14].

Equation (3) can be written in a suggestive form as[5]

$$\left.\frac{\delta T}{T}\right|_O=\left.\left(-\varphi_\chi-\frac{aH}{k}v_\chi+\tfrac{1}{4}\delta_{(\gamma)v}\right)\right|_E, \tag{4}$$

where $v_\chi\equiv v-(k/a)\chi$ and $\delta_{(\gamma)v}\equiv\delta_{(\gamma)}+4(aH/k)v$. $\delta_{(\gamma)v}$ is the same as $\delta_{(\gamma)}$ in the comoving gauge which sets the velocity variable $v\equiv 0$; in a pressureless matter the test particles follow geodesics and hence the comoving gauge is equivalent to the synchronous gauge which fixes $\alpha\equiv 0$. Thus, in this form the variables are viewed (evaluated) in mixed slicing (gauge). We can show that each of these gauge-invariant variables most closely resembles the Newtonian counterparts. $-\varphi_\chi$, δ_v, and v_χ most closely reproduce the behavior of the perturbed gravitational potential $\delta\Phi$, the perturbed density $\delta_N\equiv\delta\rho/\rho$, and the perturbed velocity δv in the Newtonian context [17,15,18].

It should be obvious from the above two equations and discussion that the actual form of the terms in the right-hand side depends very much on the gauge. It is best not to yield to the temptation of interpreting the individual terms "physically"—let alone try to fix the numerical prefactors. But if one insists on doing so, then the most natural choice is to interpret the first term $-\varphi_\chi$ as due to the gravitational redshift (we have $-\varphi_\chi=\alpha_\chi=\delta\Phi$), the second term as due to the Doppler effect, and the third as arising from the radiation field. As we shall see in the next section, even this interpretation is fraught with danger but at least the coefficient of $-\varphi_\chi$ is now unity.

III. ADIABATIC PERTURBATIONS

We consider a system with radiation (γ) and matter (m). The adiabatic condition

$$S\equiv\delta_{(m)}-\tfrac{3}{4}\delta_{(\gamma)}=0 \tag{5}$$

implies

$$\delta\equiv\frac{\delta\mu}{\mu}=\frac{1+R}{1+4R/3}\delta_{(\gamma)},\quad R\equiv\frac{3\mu_{(m)}}{4\mu_{(\gamma)}}; \tag{6}$$

thus we have[6] $\delta_{(\gamma)v}\sim\delta_v=\frac{2}{3}(k/aH)^2\varphi_\chi$, which is subdominant at large angular scales corresponding to the large-scale $k/aH\ll 1$ at E. We have[7]

$$v_\chi=-\frac{2}{3}\frac{1}{1+w}\frac{k}{aH}\varphi_\chi, \tag{7}$$

for the growing solution. Thus, adding the first two terms on the RHS of Eq. (4), we finally have

$$\left.\frac{\delta T^{(\mathrm{Ad})}}{T}\right|_O=-\frac{1+3w}{3(1+w)}\varphi_\chi=-\left.\frac{1}{3}\varphi_\chi\right|_E=\frac{1}{3}\delta\Phi, \tag{8}$$

where the second equality follows by assuming matter domination at E, so that $w=0$.[8] In our case $\delta\Phi$ does not evolve in time. We stress again that the result in Eq. (8) is valid under many conditions mentioned above, especially the ones above Eq. (3); e.g., the simple result in Eq. (8) does not hold (for example) in a model with an additional cosmological constant, and in such a case we should go back to the general form in Eq. (2).

It was often stressed in the literature that the $-1/3$ factor comes directly from the metric part in the synchronous gauge whereas it gets a contribution of -1 from the metric and 2/3 from the intrinsic temperature part in the zero-shear gauge [19,9]. The origin of such gauge-dependent interpretations can be understood simply by rewriting the RHS of Eq. (3) or Eq. (4) in the respective gauge conditions. In the zero-shear gauge, $\chi\equiv 0$, we have

$$\left.\frac{\delta T}{T}\right|_O=-\varphi_\chi+\frac{1}{4}\delta_{(\gamma)\chi}. \tag{9}$$

Notice that on a large scale the temperature part $\delta_{(\gamma)\chi}$ is *dominated* [when viewed in the comoving gauge; compare with Eq. (4)] by the metric, and does not behave like an ordinary temperature. Instead, it gives $\frac{2}{3}\varphi_\chi$, and we use $-1+\frac{2}{3}=-\frac{1}{3}$ to get the final result. This is a rigorous argument, and one should not confuse this with the heuristic one mentioned in the Introduction; except for the similar division into -1 and $\frac{2}{3}$ the origins and the interpretations are completely different. Therefore, the heuristic interpretation is not based on this zero-shear gauge analysis. The synchronous gauge coincides with the comoving gauge in the matter dominated era (MDE); thus in the comoving gauge we have

$$\left.\frac{\delta T}{T}\right|_O=(-\varphi_v+2H\chi_v)+\tfrac{1}{4}\delta_{(\gamma)v}. \tag{10}$$

For an ideal fluid with $w=$const we have $\varphi_v=[(5+3w)/(3+3w)]\varphi_\chi$ for the growing mode.[9] The temperature part now behaves like the conventionally known temperature fluctuation and thus is negligible on large angular scales compared with the potential fluctuation. Therefore, using Eq. (7) with $v_\chi=-(k/a)\chi_v$ the metric part gives $-\frac{1}{3}\varphi_\chi$ directly. For the original derivation, see [2]; see also [20].

IV. ISOCURVATURE CASE

The isocurvature condition is $\delta\mu_v\equiv 0$ under which we have $\varphi_\chi=0$ and $\varphi_v=0$. This condition implies $S=$

[5]With the wave number k appearing in the equation the variables can be regarded as Fourier transformed ones. To linear order the same equations in configuration space remain valid in Fourier space as well. Thus, we ignore specific symbols distinguishing the variables in the two spaces.

[6]See Eq. (6) in [14].

[7]See Eq. (7) in [14].

[8]Using $\delta a/a=[2/3(1+w)]\delta t/t$ the heuristic argument mentioned in the Introduction also produces a result for general w at the emission epoch E [5].

[9]This follows from Eq. (18) in [14] and the conservation property of φ_v under an adiabatic condition: $\varphi_v=C$. Or see Eqs. (50), (51) in [18].

$-\frac{3}{4}(R^{-1}+1)\delta_{(\gamma)v}$. The isocurvature initial condition is imposed early in the radiation dominated era (RDE). Einstein's equations give[10] $\dot{\varphi}_v = -H(\mu+p)^{-1}\delta p_v$ which shows that the initial isocurvature perturbation can generate φ_v. For an isocurvature mode we have $\delta p_v = -(1/3)\mu_{(m)}(1+R)^{-1}S$. Assuming that the last scattering epoch E occurred in the MDE we have

$$\varphi_v = \frac{1}{3}S\int_{\mathrm{RDE}}^{\mathrm{MDE}}\frac{dR}{(R+1)^2} = \frac{1}{3}S \tag{11}$$

at E where we used $S=$const in the large-scale limit [21]; this argument was used by Liddle and Lyth in [22]. Thus, in the MDE we have $\varphi_\chi = \frac{3}{5}\varphi_v = \frac{1}{5}S$; this shows the amount of curvature perturbation φ_χ in the MDE generated from the initial isocurvature perturbation S in the RDE. In the MDE we have $\frac{1}{4}\delta_{(\gamma)v} = -\frac{1}{3}S$. Therefore, from Eq. (4), using Eq. (7) and assuming the MDE at E, we have

$$\left.\frac{\delta T^{(\mathrm{Iso})}}{T}\right|_O = -\frac{2}{5}S\bigg|_E = -2\varphi_\chi|_E = 2\delta\Phi, \tag{12}$$

which is six times larger than the adiabatic result. For original derivations, see below Eq. (3.5) of [23] and below Eq. (5.27) of [24].

V. DISCUSSION

It is clear from Eq. (3) that we can divide the terms in different ways (which is actually what the gauge choice is doing). Correspondingly there are many different ways to reach the *same* final results in Eqs. (8), (12). In other words, we have infinitely many different ways of introducing slicing and thus viewing each variable in different gauges. While doing an actual calculation we need to choose the gauge (as we mentioned, a gauge-invariant variable is equivalent to a variable based on a certain slicing condition which fixes the gauge mode completely), but the final physical results should be the same independently of which gauge we have chosen. Our results in Eqs. (8), (12) are the final results where $\delta\Phi$ can be interpreted as the perturbed Newtonian potential which is related to the density contrast through Poisson's equation.

The pseudo Newtonian method described in the Introduction is closely related to the decomposition in Eq. (9). However, as we have shown below Eq. (9) such an interpretation is not supported by analyses in that gauge, which is true even in the context of our gauge-ready form in Eq. (3). That is, interpreting the parts as being due to the gravitational redshift and the time dilation is not justified in rigorous relativistic perturbation theory. Hence it is difficult to imagine that such a heuristic argument captures the basic physics. We believe it is yet another curious coincidence in general relativity in which a pseudo Newtonian argument does lead to the correct final result.

In a classic book by Zel'dovich and Novikov [25] we find the statement: ''However,...the gravitational shift contains the factor 1/3; it is still unclear how to interpret this coefficient classically,'' which still seems to be true.

[10] This follows from Eqs. (12), (14), (18) in [18].

ACKNOWLEDGMENTS

We thank George Efstathiou and Kandaswamy Subramanian for useful discussions. We also wish to thank Ed Bertschinger, Robert Brandenberger, and Ruth Durrer for useful comments on this work. T.P. acknowledges the hospitality of IoA, Cambridge. H.N. was supported by grant No. 2000-0-113-001-3 from the Basic Research Program of the KOSEF. J.H. was supported by Grants KRF-2000-013-DA004, 2000-015-DP0080, and 2001-041-D00269.

[1] J. Binney and S. Tremaine, *Galactic Dynamics* (Princeton University Press, Princeton, NJ, 1987), Sec. 2.2.1.
[2] R. K. Sachs and A. M. Wolfe, Astrophys. J. **147**, 73 (1967).
[3] J. A. Peacock, Mon. Not. R. Astron. Soc. **253**, 1 (1991).
[4] P. Coles and F. Lucchin, *Cosmology: The Origin and Evolution of Cosmic Structure* (Wiley, New York, 1995), Sec. 17.4.
[5] M. White and W. Hu, Astron. Astrophys. **321**, 8 (1997).
[6] M. S. Longair, *Galaxy Formation* (Springer, London, 1998), p. 335.
[7] L. Bergström and A. Goobar, *Cosmology and Particle Astrophysics* (Wiley, New York, 1999), Sec. 11.2.
[8] J. A. Peacock, *Cosmological Physics* (Cambridge University Press, Cambridge, England, 1999), p. 591.
[9] S. Weinberg, Phys. Rev. D **64**, 123511 (2001).
[10] P. J. E. Peebles, *The Large-Scale Structure of the Universe* (Princeton University Press, Princeton, NJ, 1980), Eq. (93.27).
[11] T. Padmanabhan, *Cosmology and Astrophysics Through Problems* (Cambridge University Press, Cambridge, England, 1996), problem 6.1, p. 345.
[12] J. Hwang and J. J. Hyun, Astrophys. J. **420**, 512 (1994).
[13] J. M. Bardeen, in *Cosmology and Particle Physics*, edited by L. Fang and A. Zee (Gordon and Breach, London, 1988), p. 1; J. Hwang, Astrophys. J. **375**, 443 (1991).
[14] J. Hwang and H. Noh, Phys. Rev. D **59**, 067302 (1999).
[15] J. M. Bardeen, Phys. Rev. D **22**, 1882 (1980).
[16] L. F. Abbott and R. K. Schaeffer, Astrophys. J. **308**, 546 (1986); M. Panek, Phys. Rev. D **34**, 416 (1986); R. Durrer, *ibid.* **42**, 2533 (1990).
[17] E. R. Harrison, Rev. Mod. Phys. **39**, 862 (1967); H. Nariai, Prog. Theor. Phys. **41**, 686 (1969).
[18] J. Hwang and H. Noh, Gen. Relativ. Gravit. **31**, 1131 (1999).
[19] J. R. Bond, in *Cosmology and Large Scale Structure*, edited by R. Schaeffer *et al.* (Elsevier, Amsterdam, 1996), p. 469.
[20] T. Padmanabhan, *Structure Formation in the Universe* (Cam-

bridge University Press, Cambridge, England, 1993), Sec. 6.3.
[21] H. Kodama and M. Sasaki, Int. J. Mod. Phys. A **1**, 265 (1986).
[22] A. R. Liddle and D. H. Lyth, *Cosmological Inflation and Large-Scale Structure* (Cambridge University Press, Cambridge, England, 2000), p. 160.
[23] G. Efstathiou and J. R. Bond, Mon. Not. R. Astron. Soc. **218**, 103 (1986).
[24] H. Kodama and M. Sasaki, Int. J. Mod. Phys. A **2**, 491 (1987).
[25] Ya. B. Zel'dovich and I. D. Novikov, *Relativistic Astrophysics, Vol. 2: The Structure and Evolution of the Universe* (University of Chicago Press, Chicago, 1983), p. 438.