Lost horizons

G. F. R. Ellis and T. Rothman[a)]
Applied Mathematics Department, University of Cape Town, Rondebosch 7700, South Africa

(Received 9 October 1992; accepted 2 March 1993)

Cosmological horizons play an essential role in determining the causal structure of spacetime and are of central importance in the inflationary universe scenario. We review the topic of horizons in simple language, pointing out a number of widespread misconceptions. The use of spacetime diagrams plotted in terms of proper time and proper distance coordinates helps sort out some of these difficulties. They complement the widely used conformal diagrams, which show causal relations clearly but severely distort proper distances.

I. INTRODUCTION

The concept of a horizon is familiar from daily experience as the distance on the surface of the earth beyond which it is impossible to see. The same idea appears in cosmology as the distance light travels within a fixed amount of time. Since the speed of light is the highest velocity at which signals can travel, cosmological horizons provide an even more effective limit on vision than the curvature of the earth; no information can be obtained from—and no causal contact is possible with—any region of spacetime beyond a cosmological horizon.

Because of their role in the causal structure of spacetime, horizons have always played an important role in cosmology and this importance has increased over the past decade or so with the advent of the inflationary universe scenario, which was largely designed to solve the "horizon problem." Yet, although the idea behind horizons seems straightforward, it is apparently one of those concepts that becomes accepted at some point during one's training rather than understood. In both the popular and professional literature one encounters statements like: "Inflation makes the horizon distance far larger than the observable universe." But if the horizon is the farthest distance one can see, this assertion amounts to, "Inflation makes the distance one can see far larger than the distance one can see." Many papers also talk about galaxies "leaving" and "re-entering" the horizon. According to relativity, such behavior is simply impossible.

A number of the same concerns were voiced in Rindler's seminal paper of 1956, where he wrote "···the meanings of many phrases used in discussions of horizons such as, for example, 'all particles on one side of the horizon,' 'crossing the horizon with the speed of light,' etc., evidently depend critically on the definitions of time and distance whose diversity is enormous. A statement meaningful and valid on one interpretation can be meaningless or false on another···".[1] Apparently things have not changed.

Even if one avoids definitional confusions, some paradoxes remain. For instance, we shall see that in the simplest expanding universe model, when the universe reaches an age t_0, the size of the horizon is $3ct_0$. How can the horizon have traveled a distance equivalent to three times the age of the universe, since its origin? Is the horizon traveling faster than light?

Because of such conceptual difficulties, and because the horizon issue has now found its way even to the pages of the *New Yorker* and *The Guinness Book of Records*, we feel an up-to-date discussion of horizons would be useful to physicists and students alike. In this paper we attempt such a discussion in language we hope will be comprehensible to anyone familiar with special relativity.

An unusual feature of our presentation is that we plot several spacetime diagrams in terms of real time and real distance before going to the conformal diagrams often encountered in general relativity. The approach allows one to think in terms of familiar quantities and makes apparent various features that are unclear in the conformal diagrams, thereby dispelling a number of common misconceptions. Such physical coordinates, however, require the introduction of *local light cones* to make clear that no superluminal signaling is involved anywhere. This complication brings with it a pedagogical advantage in showing how careful one must be when dealing with coordinate-dependent quantities.

Those already familiar with metrics as used in relativity and the basics of the standard cosmological model may want to skip Secs. II–IV. Conformal diagrams are introduced in Sec. XI and the proper solution to the horizon problem appears in Sec. XII. A word on event horizons is given in the Appendix. A few other technical papers that cover some of the same ground are Refs. 2–5.

II. METRICS

It is well known that every paper on relativity begins with a metric.[6] Let us therefore write down the most familiar metric:

$$ds^2=dx^2+dy^2+dz^2. \tag{1}$$

This is of course the differential form of the Pythagorean theorem. The quantity ds is usually called the *line element* and gives the infinitesimal distance between two points P, Q in a flat (Euclidean) three space, once dx, dy, and dz are known, that is, once the differences in the x, y, and z coordinates are specified. As written ds only represents a *coordinate* distance between P and Q, rather than a physical distance. The values for x,y,z are simply numbers, perhaps 1,2,3. Until we specify whether x,y,z are measured in centimeters or furlongs, one does not know what physical length ds represents. Coordinate distances change under coordinate transformations, whereas physical distances are invariant under such transformations. From now on we shall refer to physical length as *proper distance*, the term used in relativity.

Because the principles of general relativity require that the laws of physics remain unchanged under coordinate transformations, such transformations are central to the theory. If the coordinates in Eq. (1) are to be "pure

numbers"—dimensionless—we then need some method to set the scale of measurement, that is, to find proper distances. The easiest way to do this in Eq. (1) is to multiply the right-hand side by an arbitrary (constant) *scale factor*

$$ds^2=R^2(dx^2+dy^2+dz^2). \tag{2}$$

Thus, if R is doubled, ds is doubled, and so on. Setting R to a particular value in kilometers or cubits fixes the proper distance ds corresponding to specific coordinate increments. (One must be careful here. For a given R, the transformation $\bar{x}=Rx$ brings the metric back into the form $ds^2=d\bar{x}^2+d\bar{y}^2+d\bar{z}^2$, apparently "unscaling" ds; however, this conclusion assumes that R and dx are dimensionless. If, say, R is given in kilometers, the transformation in fact throws the dimensions onto $d\bar{x}$. Thus the physical scale is retained, but the "coordinate transformation" has dimensionalized the coordinates, which we wish to avoid.)

A good illustration of the role of R is the ordinary globe, which can be described by the following metric in polar coordinates:

$$ds^2=R^2(d\theta^2+\sin^2\theta\, d\phi^2), \tag{3}$$

where ϕ is the latitude and θ the longitude. In this case, R, the radius of the sphere, scales all distances on its surface (the angles themselves are necessarily dimensionless).

Turning to special relativity, the spacetime distance is of course described by the Minkowski metric

$$ds^2=-c^2dt^2+dx^2+dy^2+dz^2, \tag{4}$$

or

$$ds^2=-c^2dt^2+dr^2+r^2(d\theta^2+\sin^2\theta^2\, d\phi^2), \tag{5}$$

where we have written the same metric in both rectilinear and polar coordinates and, contrary to convention, have set $c=c$. The Minkowski metric describes a flat, but non-Euclidean spacetime. According to the standard interpretation, $ds^2<0$ represents a "timelike" interval, a displacement of a particle at $v<c$; $ds^2>0$ represents a "spacelike" interval, or an instantaneous displacement; whereas if $ds^2=0$ it represents motion at the speed of light (see, e.g., Ref. 6 for more details).

III. THE STANDARD COSMOLOGICAL MODEL

Cosmology tends to be concerned with metrics that represent expanding universes. The standard Friedmann–Lemaître–Robertson–Walker (FLRW) cosmology may be thought of as a generalization of the Minkowski metric to spaces that expand and that may have spatial curvature

$$ds^2=-c^2dt^2+R(t)^2\left(\frac{dr^2}{1-kr^2}+r^2(d\theta^2+\sin^2\theta\, d\phi^2)\right) \tag{6}$$

(see, e.g., Refs. 6,7). Note the resemblance to both the globe [Eq. (3)] and the Minkowski metric [Eq. (5)]. In these universes, the entire effect of the expansion is contained in the time-dependent scale factor $R(t)$. The parameter k, determining the spatial curvature, has been scaled to take on the values $+1$,0 or -1. A $k=0$ universe has flat spatial sections; a $k=+1$ universe is the higher-dimensional generalization of a globe and is commonly referred to as "closed;" whereas the $k=-1$ case represents a universe with negatively curved spatial surfaces, an "open" model.

What about the function $R(t)$? This depends on the particular model under consideration. Due to the high symmetry of the FLRW model, the 10 field equations of general relativity are boiled down to one

$$\left(\frac{\dot{R}}{R}\right)^2=\frac{8\pi G}{3c^2}\rho-\frac{Kc^2}{R^2}+\frac{\Lambda c^2}{3}. \tag{7}$$

Here, ρ is the *energy* density of the model and Λ is a constant, the cosmological constant. For now we assume $k=\Lambda=0$. In a "matter-dominated" universe, one filled with ordinary matter, $\rho\propto(\text{vol})^{-1}\propto R^{-3}$ since, as mentioned, doubling R doubles all length scales. Then (7) gives $R\propto t^{2/3}$. This is the so-called Einstein–de Sitter universe, the simplest cosmological model.

On the other hand, in a "radiation-dominated" model, the energy density of photons goes like $(\text{vol})^{-4/3}$, so $\rho\propto R^{-4}$ and $R\propto t^{1/2}$. These dependences will soon prove to be important. In particular, since $R(t)$ governs the distance between galaxies, in the matter-dominated case galaxies (and other particles locally at rest with respect to them) will have world lines that follow trajectories $\propto t^{2/3}$. We will use this case as representing a good approximation to the universe at recent times. By convention, we choose the origin of time so that the Big Bang ($R=0$) occurs when $t=0$.

IV. NULL RAYS

In the usual spacetime diagram for special relativity, light travels along 45° lines. Along such *null rays* the interval ds is always zero. Similarly, setting $ds=0$ in the standard model (6) gives the light trajectories for that universe. Because the FLRW metric is everywhere spherically symmetric, we can without loss of generality choose the coordinates so that a general null ray is represented as a radial null ray, with $d\theta=d\phi=0$. Then Eq. (6) gives

$$c^2dt^2=\frac{R(t)^2}{1-kr^2}\,dr^2 \tag{8}$$

for the infinitesimal proper distance corresponding to a coordinate distance dr (traversed in time dt).

Here we encounter the first place where conceptual clarity is essential. Assume that the time interval dt is fixed. Then the above expression appears to indicate that as light crosses the coordinate interval dr, the distance traveled depends on the function $R(t)$. For instance, it might seem it travels a different distance in a model where $R\propto t^{2/3}$ than in a model with $R\propto t^{1/2}$. This seems highly improbable if one believes that the speed of light is a constant.

The crucial point here is that the coordinate system of Eq. (6) is a peculiar type of coordinate system used by cosmologists, known as *comoving coordinates*. The coordinates of a galaxy, r, θ, ϕ are fixed in this system; the only thing changing is $R(t)$. To visualize comoving coordinates it is best to resort to the balloon analogy and imagine each galaxy stuck to the balloon at a *constant* r,θ,ϕ. However, the coordinate grid itself is getting bigger. If in Eq. (8) dt is fixed and $R(t)$ increases, then the coordinate distance dr (and hence, number of galaxies encountered by the light) must *decrease*

This actually makes intuitive sense. If an observer is sitting on galaxy A and receiving light signals from neigh-

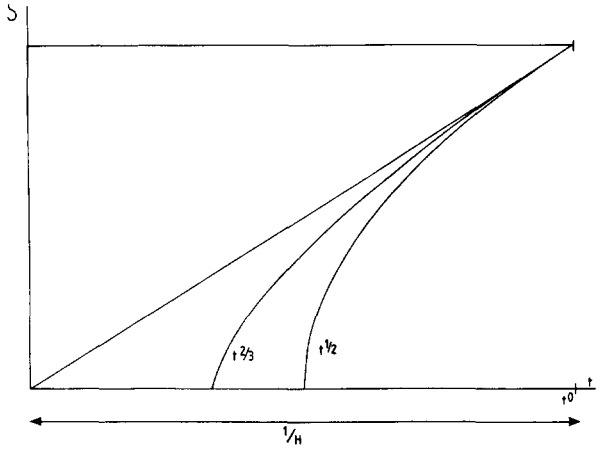

Fig. 1. One might think a $t^{2/3}$ universe expands faster than a $t^{1/2}$ universe, but the expansion rate is fixed *today* by measurements of the Hubble constant $H \equiv \dot{R}/R$. Thus the constants of proportionality are such that a $t^{1/2}$ universe is younger, and hence must be expanding faster, than a $t^{2/3}$ universe. The so-called Hubble age of the universe, $1/H$ is marked.

boring galaxies B, C, D,..., then the faster the universe is expanding, the *fewer* galaxies the observer expects to see when the light travels a given time interval dt. Since in comoving coordinates, galaxies are attached to different r's, the fewer r's an observer sees, the fewer galaxies.

As an important aside, a universe with $R \propto t^{2/3}$ is *not* expanding faster than a $R \propto t^{1/2}$ universe. With $\Lambda = k = 0$ one can easily integrate Eq. (7) for matter- and radiation-dominated models to get the age of the universe

$$t_0 = \tfrac{2}{3} H_0^{-1} \quad (matter), \tag{9}$$

$$t_0 = \tfrac{1}{2} H_0^{-1} \quad (radiation), \tag{10}$$

where

$$H_0 \equiv \dot{R}_0/R_0 = \left(\frac{8\pi G \rho_0}{3c^2}\right)^{1/2}$$

is the current value of the Hubble constant $\equiv \dot{R}/R$. This means that a $t^{1/2}$ universe is actually younger than a $t^{2/3}$ universe and hence must expand faster to reach the current value of R. The point that is usually forgotten is that we fix R and H *today*, so the curves match now. The situation is shown in Fig. 1.

V. PARTICLE HORIZONS

For a given observer at the present time t_0, the *particle horizon*, following Rindler,[1] is the surface in ordinary three-space that divides particles that have already been seen by the observer at time t_0, from particles that have not yet been seen.

To calculate the horizon distance, define the *coordinate horizon* as the coordinate distance light has traveled to an observer at t_0 since the earliest possible time t_e for emission, which in the usual case is taken to be zero (the origin of the universe). It is easily found from Eq. (8) to be

$$u = \int_0^{r_0} \frac{dr}{\sqrt{1-kr^2}} = \int_0^{t_0} \frac{cdt}{R(t)}. \tag{11}$$

Now the essential point is that there are particles (or galaxies) whose r-coordinate value is currently greater than u;

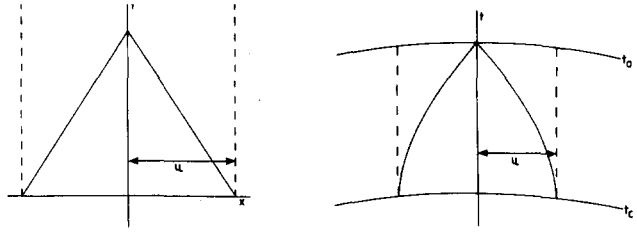

Fig. 2. In special relativity light travels along straight lines and the light cone is the familiar one. General relativity models the universe as a Riemannian manifold and the light cones will therefore be curved, as are lines of longitude on the surface of a sphere. The coordinate distance u that light travels in an interval between times t_e and t_0 is indicated. This is the coordinate horizon when t_e is the beginning of the universe. Note that this diagram does not represent spatial distances accurately.

we cannot have seen them. Those galaxies with r-coordinate value equal to u therefore separate the ones we can have seen from those we cannot.

However as before, u is just a coordinate distance. To get the proper distance to the horizon, one must again set the scale

$$h(t_0) = R(t_0)u = R(t_0) \int_0^{t_0} \frac{cdt}{R(t)}. \tag{12}$$

The quantity $h(t_0)$ is the present size of the particle horizon. It is an integral, hence a nonlocal quantity, and the upper limit of integration is *now*. This point is crucial: as defined, the horizon refers to a distance *today*, not to the position of any particle in the past. If one remembers this point, certain headaches will be avoided in what follows.

Figure 2 shows the past light cone of an observer at t_0 with the coordinate horizon indicated. The diagram is schematic and meant to resemble the usual spacetime diagram from special relativity but in curved space. It will turn out this sort of diagram is not the most useful.

At this stage, one encounters two interesting issues; one a curiosity, one a major conceptual difficulty. The first is that for a static universe $R = constant$ and h is immediately found to be $h = ct_0$, as you might expect. However, this assumes that the static universe came into existence at $t_e = 0$. For an infinitely old universe, $t_e \to -\infty$, the integral in Eq. (12) diverges at the lower limit and there is no horizon—all galaxies are visible. This is important when thinking about Olber's paradox (see, e.g., Harrison, Ref. 7). However the real universe is not static.

The second issue is a that for the standard matter and radiation models, h is easily found to be

$$h = 3ct_0 \quad (matter), \tag{13}$$

$$h = 2ct_0 \quad (radiation). \tag{14}$$

This is apparently very strange. If the speed of light is the universal constant c, how in a time t_0 can one see a distance $2ct_0$ or $3ct_0$?

The short answer is that no signal is propagating faster than light and no signal has traveled a distance $3ct_0$—a point that will become clearer as we go along. For now we note that a galaxy on the horizon is at a distance $= 3ct_0$ today; the "galaxy" was in fact on top of us when the light we see was emitted (its distance was then zero; see Fig. 3). The galaxy as it exists today is not yet visible to us.

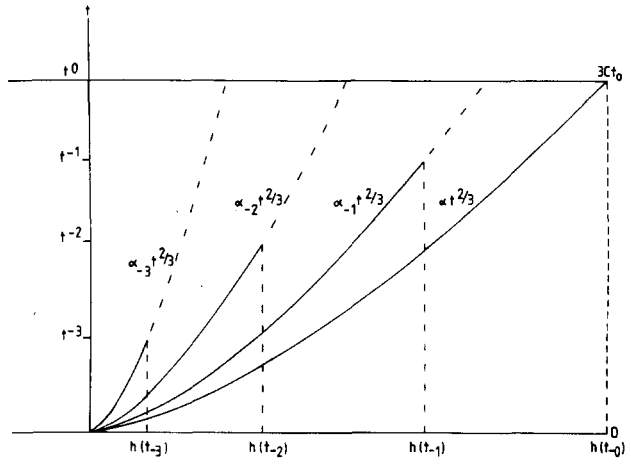

Fig. 3. Equation (15) shows that the horizon distance $h \propto R \propto t^{2/3}$. The proper distance between galaxies also increases as R, so both the world lines of galaxies and h follow curves $\propto t^{2/3}$. Indeed, the horizon may be considered as the set of galaxies that separates the seen from the unseen. Several trajectories are shown here, representing the horizons at times t_3, t_2, t_1, and t_0. The curve marked $\alpha_0 t^{2/3}$ delimits the horizon today.

On the other hand, this appears to compound the problem; if a galaxy on the horizon has traveled a distance $3ct_0$ in a time t_0, then the velocity of the galaxy must have been at least three times the speed of light! One must seize the bull by the horns and accept that this is in fact what happens.

To get a mental picture of how the horizon can recede at $2c$ or $3c$, consider the usual expanding balloon with radius R. Two galaxies are separated on the surface by an arclength S and a constant angle θ, with $S/R=\theta$. Clearly, $\dot{S}=\theta\dot{R}$. If, e.g., $\dot{R}=c/2$, then $\dot{S}>c$ when $\theta>2$ rad. The two galaxies are receding faster than c but, as we will show, no signal can be exchanged between them.

First, in order to better visualize the horizon, we plot the current horizon's time behavior. We want the proper distance [measured in a surface (t=const)] at some time t in the past, of those galaxies that become today's horizon. Since we use comoving coordinates, this is given simply by multiplying u by $R(t)$

$$h(t_0,t)=R(t)u, \tag{15}$$

where u is the coordinate horizon as defined in Eq. (11). Today's horizon is found by setting $t=t_0$. If $R \propto t^{2/3}$ then $h \propto t^{2/3}$ as well (although as a function of t_0, $h \propto t_0$). Note that because $h \propto t^{2/3}$, the horizon is behaving just like the world lines of galaxies (see Sec. III). For conceptual clarity, it is in fact useful to follow Rindler in considering the horizon to *be* the set galaxies beyond which we have not yet seen. Figure 3 shows a set of horizon curves drawn in proper time and proper distance.

Note, of course, that the horizon is constantly encompassing larger amounts of the universe—more galaxies are always becoming visible as we view the universe at later and later times. Figure 3 also makes clear that once a galaxy falls within the horizon it remains within the horizon. This will become more obvious in Sec. XI, but we can already see that statements referring to galaxies entering and then leaving the horizon are false.

We have now plotted the distance to the horizon as a function of time. We will plot the velocity of specific galaxies in Sec. VII, showing they are superluminal, but Fig. 3 confirms that at time t_0 the horizon is at $3ct_0$, meaning the average recessional velocity is $3c$. However, this motion has the character of a phase velocity. The question is whether any superluminal signaling is involved. To show that this is not the case, we examine the behavior of the light rays themselves.

VI. THE PAST NULL CONE

Consider the length

$$l(t_e)=R(t_e)\int_{t_e}^{t_0}\frac{cdt}{R(t)}. \tag{16}$$

This represents the *proper* distance l corresponding to a time t_e when a light signal was emitted such that it reaches us today at t_0. In other words, $l(t_e)$ for all $t_e<t_0$ is the locus of points that lie on our past light cone. (Just as in special relativity the past light cone is the locus of points such that all signals emitted reach us now.) For $R \propto t^{2/3}$ we easily find

$$l(t_e)=3c(t_e^{2/3}t_0^{1/3}-t_e). \tag{17}$$

This is a striking result: $l(t_e)$ is not a monotonic function of t_e. Setting $dl/dt_e=0$ shows that the maximum l occurs at $t_e/t_0=8/27$ and $l_{\max}=4ct_0/9$. Figure 4(a) shows $l(t_e)$ along with $h(t_0,t)$. Note again the proper time and distance coordinates: these are the physical times and distances as would be measured directly by local experiments.

How do we explain the shape of the past light onion? Here is an example of gravitational lensing: light rays emitted by a distant enough galaxy in the past initially spread out, but are then refocused by the matter in the universe and are reconverging by the time they reach us. Conversely as we follow our light cone back into the past, the gravitation of the matter it encloses causes refocusing, so it reaches a maximum size (the "equator") and then contracts.

From Fig. 4(a) we see that no light ray received now ever reaches us from anywhere near the horizon distance; thus no signal conveying information to us travels faster than light. One might be initially perplexed that the horizon is always outside the past light cone in the diagram. This is as it should be; the horizon separates particles that have already been seen from those that have not yet been seen. Anything to the left of the horizon has been observed by the current time; anything to the right has not. To make this clearer, in Fig. 4(b) we plot several world lines of galaxies that formed the horizon at earlier times (as in Fig. 3). Each of these earlier horizons has intersected our present past light cone. Just as in special relativity, that means the corresponding particle will have already been seen by us. The current horizon curve is tangent to the past light cone at $t=0$.

VII. THE SPEED OF LIGHT SPHERE

Sometimes people have the impression that the horizon is the distance at which a galaxy is receding at the speed of light [see, e.g., the *Guiness Book of Records*, 1992 edition, p. 6 ("Remotest object")]. It is easy to see that this is not the case. As before, when $k=0$ the proper distance to a galaxy (expressed in terms of comoving coordinates) is

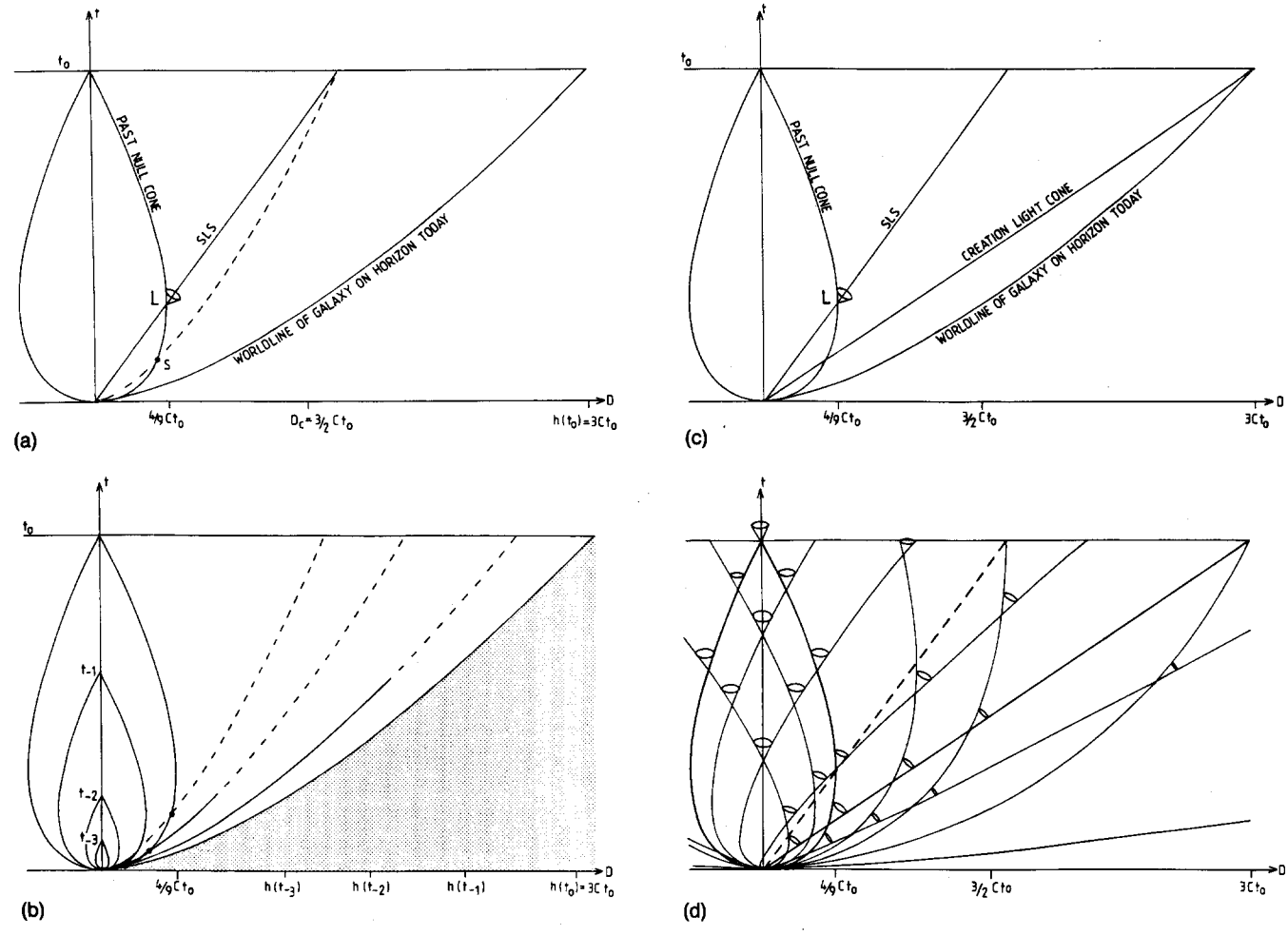

Fig. 4. (a) Today's past light cone [Eq. (17)] is shown as well as the corresponding horizon curve. (a)–(d) represent spatial distances accurately. Note that today's value of the horizon is $3ct_0$ as found in Sec. III. But no light ray from our past light cone starts at anywhere near this distance; indeed the maximum distance from which we receive light is $4ct_0/9$, showing that nothing conveying information to us has traveled faster than light. Also shown is the speed-of-light-sphere, SLS, on which objects are receding from us at the speed of light. It corresponds to a distance $D_c=3ct/2$. The point L marks the intersection of the past light cone with the SLS. The point S marks the position at which a galaxy receding at the speed of light *today* was actually seen by us. (b) Any galaxy to the right of today's horizon (the shaded region) has not yet been seen by us. Any galaxy to the left of today's horizon has followed a world line that has already intersected our past null cone, as can be seen by examining trajectories of previous horizons (cf. Fig. 3). These horizons correspond to the past null cones drawn. The points where the world lines intersect today's past null cone mark where we have seen galaxies on those world lines. (c) As in (a), but with the creation light cone. The creation light cone gives the horizon when it reaches $t=t_0$. (d) The global light-cone structure of the FLRW spacetime with local light cones drawn in. This shows that in proper-distance coordinates the speed of light varies considerably from one place to another.

$$D=R(t)r. \tag{18}$$

Since r is constant in time for a galaxy, the time derivative is

$$v=\dot{D}=\dot{R}r=\frac{\dot{R}}{R}D, \tag{19}$$

which is Hubble's law (the recessional velocity of a galaxy is proportional to distance.)

Now, setting $v=c$ gives a distance $D_c=3ct_0/2$ for the matter-dominated case. Another surprising result. The horizon is *not* the distance at which objects are moving away with velocity c. The horizon is twice this distance; the matter there is moving away from us at twice the speed of light. We plot the speed-of-light sphere (SLS) on Fig. 4(a).

There is another way to interpret this result. The Hubble radius, as defined by astronomers, is $c/H\equiv cR/\dot{R}$. Note that $R/\dot{R}=3t/2$. Thus the *Hubble radius of the universe corresponds to the SLS, not to the horizon* [in Harrison's recent paper on this topic (Ref. 3) he refers to the SLS as the Hubble sphere]. We return to this important point in Sec. XIII.

It is true, however, that the horizon is the surface of infinite redshift, as can be seen from the cosmological redshift formula

$$z=\frac{R(t_0)}{R(t_e)}-1. \tag{20}$$

Thus if $R(t_e)\to 0$ at $t_e\to 0$, then the redshift $z\to\infty$. Consequently infinite redshift is not the same as motion away from us at the speed of light.

It is important to keep in mind that, as with the horizon, a galaxy currently receding at the speed of light was actually seen by us at a much earlier time (when it was moving away from us much faster than the speed of light). This

time and position is marked S on Fig. 4(a): it is the intersection of our past null cone with the worldline of the galaxy that now lies on the SLS.

With this in mind, the redshift formula gives immediately another striking result. Solving Eqs. (17) and (19) for the intersection of l and the SLS, one quickly finds that the SLS intersects the past null cone at its maximum (this seems to be a coincidence). The maximum occurred at $t_e/t_0=8/27$, at which point $R(t_e)/R(t_0)=(8/27)^{2/3}=4/9$. Equation (20) then gives $z=1.25$ at the SLS. In other words, galaxies at point L on Fig. 4(a) emitted the radiation we now see when they were receding from us at the speed of light. Such galaxies have a redshift of only 1.25—a far cry from infinity. This means that all the objects we see with redshifts $z>1.25$ (which include many quasars) gave off the light we observe when they were moving away from us faster than c. Indeed, according to Eq. (7), the velocity of all matter was infinite at the Big Bang; it then gradually slows to subluminal velocities.

What is the present position of the matter that we observe at the exceptional redshift value of $z=1.25$? It is in fact just a distance ct_0 away from us today. The *average* recessional velocity of this matter, since the Big Bang, is precisely c! (And c is also precisely the speed of the matter when it emitted the light by which we see it!) Contrast this with the matter at point S on Fig. 4(a). This matter is *today* moving away from us at the speed of light. The time of emission is easily calculated to be $(1/8)t_0$. The redshift of this matter is $z=3$, and at t_e the velocity was $2c$, which is greater than c (as it must be).

We also take this opportunity to mention that because $z=1.25$ is at the maximum of the past light cone, where gravitational refocusing of light begins, this redshift is also the position of minimum angular diameter of galaxies (remember we are assuming the Einstein–de Sitter case: $p=\Lambda=k=0$).

VIII. LOCAL LIGHT CONES

One might doubt the statement that no signal is propagating faster than light is true at the point marked L on Fig. 4(c)–the point where the SLS intersects the past light cone. At L a galaxy is seen by us at t_0 to be receding at the speed of light. Does an observer sitting at L measure other matter to be receding faster than c? No! To understand this one must calculate the behavior of light in the vicinity of L; we must determine the local light cone. We show how to this now, although readers not interested in the details can skip to Sec. IX.

Recall first that the metric Eq. (6) is comoving, but our graphs are plotted in real (proper) time and real (proper) distance D. To see what the local light cones look like on such graphs we transform from comoving coordinates $\{t,r,\theta,\phi\}$ to proper distance coordinates $\{t,D,\theta,\phi\}$. From Eq. (18), when $k=0$

$$dD=Rdr+rdR \tag{21}$$

or

$$dr=\frac{dD}{R}-\frac{D}{R^2}\dot{R}dt. \tag{22}$$

This gives

$$R^2dr^2=dD^2-D^2\frac{\dot{R}^2}{R^2}dt^2-2D\frac{\dot{R}}{R}dDdt. \tag{23}$$

The radial part of the metric Eq. (6) now becomes

$$ds^2=\left(-1+D^2\frac{\dot{R}^2}{R^2}\right)c^2dt^2-2D\frac{\dot{R}}{R}dDdt+dD^2. \tag{24}$$

For null rays, $ds=0$, as always. Dividing both sides by dt^2 gives a quadratic equation in dD^2/dt^2 which is easily solved to yield

$$\frac{dD_\pm}{dt}=D\frac{\dot{R}}{R}\pm c. \tag{25}$$

The quantity dD/dt is the tilt (inverse slope) of light rays at an arbitrary point on the graphs; in other words, it gives the local light cone at any point.

Since $D\dot{R}/R$, Hubble's law, represents the recessional velocity of matter, we see that the relative velocity of light and matter is always c. Specifically, at the SLS, Eq. (19) gives $dD/dt=D\dot{R}/R=c$. Inserting this into Eq. (25) gives

$$\frac{dD_-}{dt}=0;\quad \frac{dD_+}{dt}=2c \tag{26}$$

for the local light cones. That is, one null ray is vertical while the other has tilt $2c$. This will be true anywhere along the SLS and is independent of equation of state. Now, at the SLS the velocity of matter $\dot{D}_m$ was by definition c. The tilt of local light cone there was $\{0,2\}$, telling us that the velocity of light relative to matter is still c, as required by relativity. Nothing is propagating faster than local velocity of light. The local light cone is drawn in Fig. 4(c).

Using Equations (25) and (27) below, one can find the light-cone structure not just at the SLS but over all spacetime; the cones are everywhere tangent to the light rays through each point. The result is shown in Fig. 4(d). [To aid plotting, note that dD_+/dt in Eq. (25) is always positive, but dD_-/dt changes sign at the SLS. The SLS thus represents the turning points of these trajectories.]

IX. THE CREATION LIGHT CONE

The peculiarity that the velocity of light at L is $2c$ stems from the coordinates we have chosen, which extend from the origin over all spacetime. General relativity attempts to make spacetime resemble the flat Minkowski space of special relativity—where the speed of light is always c—but it can only do this locally, not globally. Locally, one can always find a coordinate system so that the velocity of light is c. The details of how to do so are a large part of a general-relativity course. Nevertheless, we have already shown that locally the relative velocity of matter and light is c, so one should find the following result easier to accept. Equation (25) integrates to

$$D_\pm=\pm R(t)\int_{t_1}^{t}\frac{cdt}{R(t)}, \tag{27}$$

where we have used the definition of r and where t_1 is an integration constant. By convention the $(+)$ indicates null rays outgoing from t_1, while $(-)$ indicates null rays ingoing to t_1 [although $t>0$ at all points in the spacetime, to plot geodesics that lie outside the horizon, it is necessary to let t_1 go negative in Eq. (27).] With $t_1=t_0$, we have

$$D_-(t)=-R(t)\int_{t_0}^{t}\frac{cdt}{R(t)}=+R(t)\int_{t}^{t_0}\frac{cdt}{R(t)}, \quad (28)$$

which is our previous result for the past null cone [cf. Eq. (16)].

On the other hand, the light cone spreading out from $t_0=0$ to the future is given by $t_1=0$: then

$$D_+(t)=R(t)\int_0^t\frac{cdt}{R(t)}. \quad (29)$$

This is termed the creation *light cone*. For a matter-dominated universe, $D(t)=3ct$. The creation light cone is plotted on Fig. 4(c) and looks nothing like the past light cone. Here is perhaps the hardest result to understand: the velocity of light on this trajectory, measured in terms of proper distance from the origin divided by proper time, is precisely $3c$ at all times. No getting around it. The point is that the speed of light is a constant $=300\,000$ km/s *relative to matter*. But because the matter is itself receding from us, the speed of light relative to us must be greater than c, although no causal violation is involved. Thinking again of the expanding balloon with the horizon receding faster than c, the matter velocity at $D=3ct$ is $\dot{D}=D\dot{R}/R=2c$, so as always the relative velocity of matter and light is c.

The importance of the creation light cone is that it gives the particle horizon, when evaluated at $t=t_0$. We can first receive light from a distant galaxy when it enters our horizon. Because the FLRW universe is homogeneous, that is the same instant when light from our galaxy is first visible to the distant galaxy. Figure 4(c) shows this takes place when that galaxy lies on our creation light cone and vice versa.

X. THE VISUAL HORIZON

In order to relate the horizon idea to real observations, we must introduce one other type of horizon, the *visual horizon*.[2] The visual horizon is merely the distance from us of light which was emitted at the surface of last scattering, at a time t_d (for decoupling) and a redshift $z\simeq 1000$. This is the time at which matter and radiation decoupled in the early universe. At higher redshifts and earlier times the universe is opaque to electromagnetic radiation; the visual horizon thus really does define the maximum distance one can see. Evaluated at t_d it gives the distance at which the cosmic microwave background radiation was emitted. Today it gives the present position of the particles that emitted that radiation; these are the furthest objects we can ever have seen (by any form of electromagnetic radiation).

For the above value of z, with $R\propto t^{2/3}$, Eq. (20) gives

$$t_d=(1000)^{-3/2}t_0. \quad (30)$$

The proper distance to the light cone at that time is given by Eq. (17) as

$$h_{vh}(t_d)=3c(t_d^{2/3}t_0^{1/3}-t_d)\cong\frac{3ct_0}{1000}(0.97). \quad (31)$$

This was the position of the emitting matter at t_d. Since the distance to the horizon has increased by a factor of 1000 since then (Sec. V), its position is now

$$h_{vh}(t_0)=0.97(3ct_0)=0.97h(t_0), \quad (32)$$

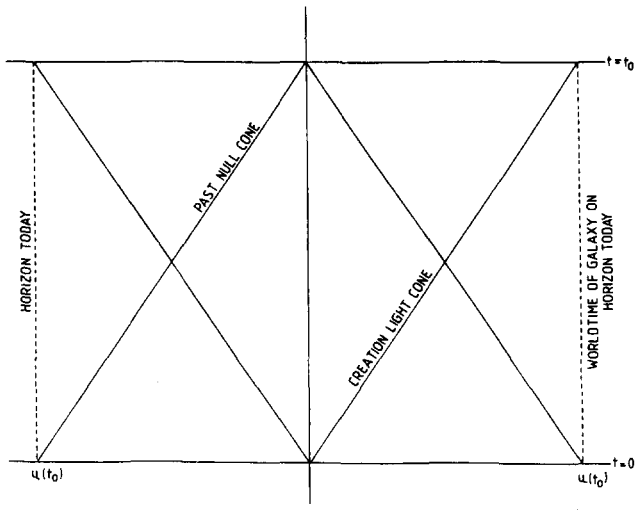

Fig. 5. The conformal version of (c) (but with the SLS omitted). All the light trajectories have been stretched into straight lines to make the figure resemble the spacetime diagrams of special relativity. The penalty one pays is that spatial distances are severely distorted. In particular, at $t=0$ (where space time becomes singular), what is represented as a single point on the previous diagram is mapped to an infinite plane.

where $h(t_0)$ is, as usual, the distance to the particle horizon. However, as we have discussed, we cannot see to this set of events, precisely because this distance is evaluated today. Recent statements that "the primordial lumps in the microwave background found by COBE represent the beginnings of galaxies" are not verifiable through astronomical observation, in the sense that any given lump observed by COBE at the visual horizon evolved into something that today lies far beyond what is now visible (Fig. 6).

The importance of the visual horizon will become clear in Sec. XII. For the moment we note the following: the *maximum distance to which we can see* in the universe is $l_{\max}=4ct_0/9=(8/29)(1/H_0)$, as was already established (Sec. VI). *The distance we see by observing the microwave background radiation* is much less that this; it is $h_{vh}(t_d)$ [given by Eq. (31)], only about 10^7 light years. Further, when this matter emitted that light, its speed of motion away from us was (cf. Sec. VII)

$$v=\frac{\dot{R}}{R}(t_d)D_d=\frac{2}{3t_0}(1000)^{3/2}\frac{3ct_0}{1000}(0.97)$$
$$=1.94c1000^{1/2}=61.3c. \quad (33)$$

This is the fastest moving matter we can see in the universe! However at the present day that matter is moving away from us at $2c$. Note that in both cases, the speed of motion of that matter measured in a local comoving frame is precisely zero (for that is the definition of such a frame).

XI. CONFORMAL DIAGRAMS

So far we have plotted the results in proper time and proper distance. This method has the advantage of using familiar quantities and it shows the real distances light has traveled in a time dt, but it has several disadvantages. Chief among them is that the speed of light is no longer the local speed of light. Further, all the past null cones and horizon

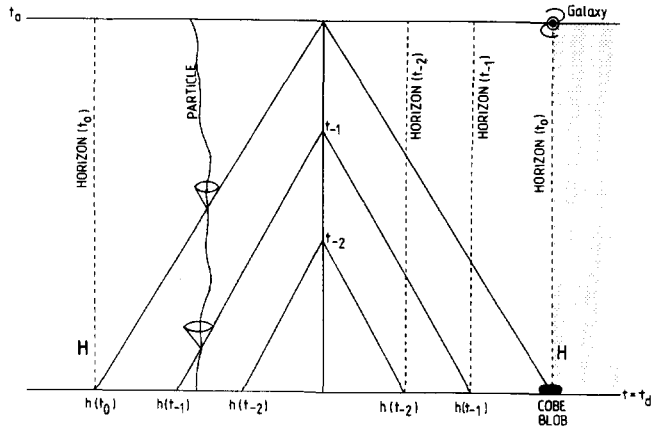

Fig. 6. A conformal diagram illustrating the increase of the horizon with time [cf. Fig. 4(b)]. The world lines of galaxies are vertical. This shows clearly that a lump observed by COBE at the visual horizon has evolved today into something that is not currently visible. Also shown is the world line of an arbitrary particle not at rest with respect to galaxies. Since its trajectory must be bounded by 45° lines, once it enters the horizon, it remains within it.

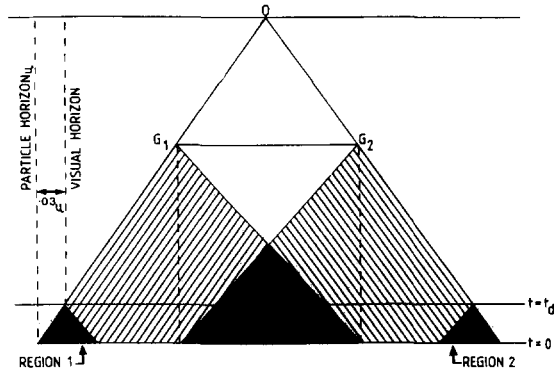

Fig. 7. Two galaxies G_1 and G_2 that we see in opposite directions in the sky at the moment they are first able to causally interact. From basic geometry it is easy to calculate that the distance between them is $2u/3$. The diagram also shows the relationship between the visual horizon (the furthest distance we can *see* today) and the particle horizon. One sees that the difference between them $\sim 0.03u(t_0)$ was the size of u at t_d. Therefore regions 1 and 2 shown at the bottom were separated by about 60 horizon distances at the time of decoupling. It was thus impossible for them to causally interact. This is the horizon problem.

lines converge at $t=D=0$ and the details there are impossible to discern. Both problems can be alleviated by rewriting the FLRW metric Eq. (6) as

$$ds^2=R^2[-d\eta^2+d\tilde{r}^2+\tilde{r}^2(d\theta^2+\sin^2\theta\, d\phi^2)], \qquad (34)$$

where we have used $\tilde{r}$ to represent a comoving coordinate system where the $1-kr^2$ does not appear. It is clear that such a transformation can be made in the $k=0$ case. Then $\tilde{r}=r$, $R=R(t)$, and $d\eta(t)\equiv cdt/R$; one can show[8] that it is possible to make a similar transformation for all the FRLW models.

This *conformal transformation* has changed the FLRW model into a flat Minkowski metric[5] except for the conformal factor $R^2(x^i)$. For obvious reasons η is termed *conformal time.*

Clearly, in this coordinate system null rays ($ds=0$) travel along 45° lines, as they do in special relativity. This is a great simplification. Under the conformal transformation Eq. (34), Fig. 4(c) becomes the *conformal diagram* Fig. 5. Note the time coordinate is

$$\eta(t)=\int_0^t \frac{cdt}{R(t)} \qquad (35)$$

and in the $k=0$ case, the space coordinate is comoving coordinate value r ($\tilde{r}=r$). Surfaces of homogeneity (e.g., constant density surfaces) are horizontal on such diagrams, and world lines of galaxies are vertical.

One pays a penalty for using conformal diagrams: it completely hides true distances (correctly represented in the previous coordinates and diagrams). The proper distance between "galaxies" at the Big Bang singularity $t=0$ is zero. But because on the conformal diagram the distance to a galaxy is given in coordinate distance, it remains constant, fixed at "today's" value. The physical distances near $t=0$ are thus severely distorted, appearing much larger than they actually are; and this distortion becomes infinitely large as $t\to 0$. Nevertheless, conformal diagrams are essential in understanding fully the nature of horizons. They also make it much easier to see what happens in inflation, and so the following discussion will be based exclusively on them.

Before turning to inflation, consider Fig. 6. We see clearly how a galaxy world line H separates those that can have been seen by us, at time t_0, from those that cannot have been seen by us. We also see how our creation light cone intersects the particle horizon today, just as we cross the creation light cone of those galaxies that comprise the horizon. This diagram also makes it easy to see how the horizon increases with time. [This behavior is in sharp distinction to MaCallum's "reference horizon" (private communication), which is the horizon beyond which a published work is no longer cited. The reference horizon is a rapidly decreasing function of time, and has a current value measured at approximately 6 months.] Further, since on conformal diagrams world lines of particles are bounded by 45° lines, it is obvious that once any particle enters our horizon it remains within it. Figure 6 also clarifies the claim made in Sec. X that a blob seen by the COBE satellite develops into a structure that is invisible today. Figure 7 shows the relation between particle horizon and the visual horizon.

Note particularly that these diagrams represent exactly the same situation as shown in the previous diagrams, but using different coordinates. The proper coordinate diagrams are transformed to these ones by magnifying the spatial length scale by an amount that gets larger and larger at earlier times, in such a way that galaxy world lines become vertical and light rays travel at $\pm 45°$.

XII. WHAT DOES INFLATION DO

Inflation was designed largely to solve the "horizon problem." The horizon problem is most easily phrased in terms of the microwave background. The COBE results showed that the temperature of the background is uniform to about 1 part in 10^6. The question is, how did the background become so isotropic?

To be more precise, consider Fig. 7. An observer O at t_0 sees a highly uniform background originating at t_d, the

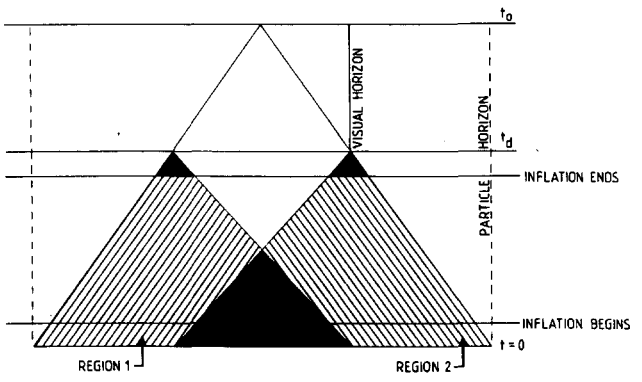

Fig. 8. Inflation solves the horizon problem by putting the universe through a period of exponential expansion at a very early time before t_d. (The diagram is unchanged after t_d; cf. Fig. 7.) The singularity at $t=0$ is thus moved exponentially far from t_d (in terms of these conformal coordinates) allowing significant overlap of the past light cones of regions 1 and 2 (the dark shaded region at the bottom). Nonetheless, the overlap is not total unless the universe is spatially closed (Sec. XIII).

decoupling time. Yet, as can easily be calculated from the figure, the two galaxies marked G_1 and G_2 that the observer sees in opposite directions could not have interacted with each other if they are separated by more than $2u/3$, two-thirds of a coordinate horizon. This corresponds to an angular separation of 60° in the sky at present. In other words, regions on opposite sides of the sky are just coming into causal contact now.

From Eq. (32) we found that $h_{vh}=0.97h(t_0)$. Both scale as R, so at the time of decoupling, t_d this remains true (i.e., the coordinate horizons always have the same value.) Referring to Fig. 7, we see that $u(t_0)-u_v(t_0)\approx 0.03u\approx 1/30u$ was in fact the horizon distance at decoupling. Thus regions 1 and 2 on the diagram were separated by about 60 horizon distances when the microwave background was created.

This being the case, causal interaction between various regions of the sky was impossible. How then, did the microwave background become so uniform? This is the horizon problem. One answer is that the universe was uniform from the outset. Another answer is that inflation made it that way.

According to the inflationary scenario, at about 10^{-35} s after the Big Bang (depending on the model) the universe underwent a brief period of exponential expansion, during which

$$R\propto e^{Ht}, \tag{36}$$

where $H=\dot{R}/R=$constant. Note that this is equivalent to $\dot{R}/R=\Lambda c^2/3$ in Eq. (7).

Because inflation took place at very early times, it affects Fig. 7 only near $t=0$; it certainly leaves it unaffected after t_d. Recalling that conformal time $\eta(t)$ squeezes an infinite amount of time onto a finite sheet of paper, the effect of inflation is to give Fig. 8. We see that inflation allows the past histories of two causally disjoint regions to significantly overlap. Consequently they may interact and make the microwave background uniform (though inflation does not necessarily specify the form of the interaction).

That is all there is to it.

XIII. WHAT DOES INFLATION NOT DO

Frequently one hears that inflation makes the horizon far larger than the observable universe. We are now in a position to understand what this statement means. Due to the exponential increase of R in Eq. (36), the *particle* horizon is made exponentially large. But the *visual* horizon, which relates to the much later time of t_d, remains exactly where it was. Thus we *see* exactly as far as we do in the standard model.

One also frequently encounters statements in the literature about galaxies leaving the horizon, then reentering the horizon at a later time. We have already explained that this is impossible. Upon closer examination, what such statements refer to is the following. Before inflation begins, a comoving length scale, λ will be getting larger with R. At the same time $\dot{R}/R$ is getting smaller as the expansion of the universe slows. Let us assume that λ represents the distance to a visible quark—the quark lies within our horizon. When inflation begins, λ gets exponentially larger while $H\equiv\dot{R}/R$ becomes constant. Thus, the distance to the quark becomes much larger than c/H, or $\lambda H/c\gg 1$. The quark leaves the horizon. After inflation, H becomes much smaller and eventually $\lambda H/c$ becomes $\ll 1$ and the quark (now part of a galaxy) reenters the horizon.

We see, however, that in this discussion c/H has been tacitly equated with the horizon distance. But we have already shown in Sec. VII that c/H is the speed-of-light sphere. Therefore this discussion has nothing to do with horizons. The same papers usually refer to c/H as the limit of causal interaction, but as we have already detailed, that distance is given by the horizon, not by the SLS.

XIV. CLOSED SPACE SECTIONS

We complete our survey by pointing out a feature of inflation that seems to have gone largely unnoticed in the literature: There is a large difference between inflation in a spatially open ($k=-1$) model and a spatially closed ($k=+1$) model.

In a $k=-1$ or $k=0$ standard model, the spatial extent of the universe is truly infinite and therefore the model contains an infinite amount of matter. The horizon, then, limits our causal contact to a vanishingly small fraction of the matter in the universe. This remains true even under inflation; although the particle horizon is made exponentially larger, it is still not infinite. Thus new information can enter the horizon in the future and it is always possible that some large-scale gradient in conditions will eventually make the background radiation anisotropic (direction dependent).

In a closed ($k=+1$) model with vanishing cosmological constant, this is far less likely to happen. Due to the curvature of space we see a finite fraction of all there is. [The precise value of this fraction depends on the famous density parameter $\Omega\equiv(\rho)/(\rho_{\rm crit})$, where $\rho_{\rm crit}$ is the value of ρ in Eq. (7) when $k=\Lambda=0$]. For instance, one can show that the horizon encompasses everything in a closed model just at the moment of final collapse—you have seen right around the universe.

However, as discussed above, if there is a nonzero cosmological constant in Eq. (7) (or equivalently a potential dominated scalar field), inflation results, and we actually see around (or at least make causal contact) before the end of the universe. To see how this works, consider Eq. (11)

with $k=+1$ and $R=R_1e^{Ht}$, where R_1 is the scale factor at the beginning of inflation. Then Eq. (11) integrates to

$$\arccos(u)=\frac{ce^{-Ht_1}}{HR_1}(1-e^{-H\Delta t}), \tag{37}$$

where t_1 is the time that inflation begins and Δt is the duration of the inflationary period considered. (Here is a good example of a model in which the horizon is definitely not the SLS: The SLS$=c/H$ is constant, but the horizon is always increasing.)

Now, the maximum value of $\arccos(u)$ needed to see around the universe is π and t_1 can be set to zero. Then we have

$$\pi=\frac{c}{HR_1}(1-e^{-H\Delta t}) \tag{38}$$

determines the time Δt needed to see right round the universe.

In the FLRW model without inflation, recalling that $R\propto \textit{temperature}^{-1}$, one can easily calculate that at inflationary temperatures $T\sim 10^{28}$ K, the "radius" of the universe was $R\sim 1$ cm, while c/H was smaller by a factor of $\sim 10^{25}$. This surprising result is actually another statement of the horizon problem. For the FLRW model $c/H\sim$*horizon*: as $t\to 0$, (horizon)/(scale factor)$\to 0$ so at early enough times no particles could communicate. Thus $c/HR\sim 10^{-25}$ in the standard model and it is impossible to see a coordinate distance π.

The situation is different if there was an inflationary era. One generally fixes R at today's value. That means if inflation took place, and increased R exponentially, then R_1 must have been exponentially *smaller* than in the standard model at the equivalent time, so $c/HR_1\gg 1$. (This is another way of stating how inflation solves the horizon problem.) In that case, Eq. (38) shows that once inflation has started, it is possible to see around the universe each time $\pi\sim H\Delta t$, or $1/\pi$ times the number of e-foldings during inflation. The upshot is that inflation can isotropize the universe much more effectively in the $k=1$ inflationary model because it allows each particle to be in causal contact with *all* the other matter in the universe many times over (you can "see" round the universe many times), which is not possible in the open or flat inflationary models. The particle horizon is thus broken at an early time in inflationary universes with closed space sections.

XV. CONCLUSION

We see that because of the curvature of spacetime, when we plot the history of the expanding universe in proper distance and proper time coordinates some surprising facts emerge. We do not see out to a Hubble radius; the microwave background radiation was emitted a very small distance from our past world line; the matter emitting that radiation was moving away from us at 60 times the speed of light at the time of emission. Further, the light forming our creation light cone traveled away from us at an effective speed of three times the speed of light, explaining why the particle horizon today is at a distance corresponding to three times the age of the universe. The galaxies forming that horizon have been moving away from us much faster than the speed of light, and are presently moving away from us at twice the speed of light. All these features follow by straightforward calculation of proper distances in a FLRW universe. The usual conformal diagrams make causal properties very clear, but hide the real spatial distances involved.

On the other hand the causal diagrams make quite clear that the particle horizon is an absolute limit on communication and the visual horizon an absolute limit on observation; once matter has entered one of these horizons, it cannot leave it. Inflation moves the particle horizon out to an exponentially large distance while leaving the visual horizon fixed. In doing so it allows causal contact of those events where the microwave background radiation was emitted, thus in principle solving the horizon problem. However a complete solution is only attained in a $k=+1$ universe (or in universes where the spatial sections are closed for topological reasons, see Ref. 6); in this case only, all the matter in the universe is in causal contact at early times, because the particle horizon ceases to exist at an early stage in the inflationary era.

ACKNOWLEDGMENTS

We thank the FRD (South Africa) and MURST (Italy) for financial support. We thank Bruce Bassett for the substantial amount of time he spent in generating the figures.

APPENDIX A: EVENT HORIZONS

For a discussion of cosmology and inflation, the most important horizon is the particle horizon. One often encounters in the literature, however, another horizon that may be more familiar because of its association with black holes. This is termed the *event horizon* and we discuss it briefly to distinguish it from the particle horizon.

The particle horizon is the distance of particles beyond which an observer cannot see at the current time. The event horizon for an observer is defined as the distance beyond which this observer will *never* see. Specifically

$$h_{\rm ev}\equiv R(t_0)\int_{t_0}^{\infty}\frac{cdt}{R(t)}. \tag{A1}$$

Thus it is the limit of the past light cone of a point on our world line as time goes to infinity; its physical size is being evaluated at time t_0.

The ever-expanding FLRW models have no event horizon—given enough time an observer will see everything. In the case of $k=+1$ models this is not the case; for these models the upper limit of integration is usually taken to be the end of the universe, leading to a finite value for $h_{\rm ev}$.

A particular model with an event horizon is the *De Sitter* universe, in which the scale factor increases exponentially: $R\propto e^{Ht}$, where H, the Hubble constant, is assumed to be constant. (The inflationary period is one in which the universe goes through a De Sitter phase.) One can easily verify that $h_{\rm ev}=c/H$ for a De Sitter universe. This is the farthest spatial distance observers at t_0 can ever hope to see to, if they can observe to the infinite future of time.

The upper limit of ∞ makes the definition of the event horizon somewhat metaphysical. If one imagines getting a grant from the World Scientific Foundation for an astronomy program that will last 100 000 yr, this is still a negligible amount of time compared to infinity. In practice, our point of observation ("here and now") does not move far into the future along our world line as we make our obser-

vations, and the current past light cone effectively defines what will be visible in the foreseeable future. Thus *in practice in cosmology the event horizon is given by nothing other than our past light cone.* This is what separates the events with which we have had causal contact from those we have not.

This problem also arises in connection with black holes. A black hole is defined by its event horizon; processes within the event horizon will never be visible to outside observers. However, its definition requires an integration to $t=\infty$. Although many researchers argue that this is not really necessary in practice, for physically speaking black holes form in a finite time, mathematically no one has succeeded in defining black holes with a finite upper limit of integration in a universe that lasts forever.

[a]Current address: Harvard-Smithsonian Center for Astrophysics, 60 Garden St., Cambridge, MA 02138.

[1]W. Rindler, "Visual horizons and world models," Mon. Not. R. Astron. Soc. **116**, 662–677 (1956).
[2]G. F. R. Ellis and W. Stoeger, "Horizons in inflationary universes," Class. Quant. Grav. **5**, 207–220 (1988).
[3]E. R. Harrison, "Hubble spheres and particle horizons," Astrophys. J. **383**, 60–65 (1991).
[4]P. T. Landsberg and D. A. Evans, *Mathematical Cosmology* (Oxford University Press, New York, 1977), Sec. 11.6.
[5]Ø. Grøn, "Repulsive gravitation and inflationary universe models," Am. J. Phys. **54**, 46–52 (1986).
[6]G. F. R. Ellis and R. M. Williams, *Flat and Curved Space-Times* (Oxford University, New York, 1988).
[7]E. R. Harrison, *Cosmology* (Cambridge University, Cambridge, MA, 1981).
[8]H. Stephani, *General Relativity* (Cambridge University, Cambridge, MA, 1982).

Square wheel

Nelson H. Klein
Department of Science and Technology, Bucks County Community College, Newtown, Pennsylvania 18940

(Received 16 November 1992; accepted 27 February 1993)

A solution to the paradoxical problem of a square wheel is presented. Using kinematics, it is shown that the correct roadbed for a square object rolling without slipping is a series of inverted catenaries. The dynamics of the square are revealed by the conservation of energy method. Remarkably, the square is shown to be capable of winning a downhill race against a sphere on a parallel inclined plane.

I. INTRODUCTION

Recently I was browsing through the *Exploratorium Cookbook*,[1] a compendium of museum quality demonstrations, and came across an entry in the mathematics section entitled Square Wheels. The article described the construction of an apparatus consisting of two squares, connected by an axle, that would roll smoothly across an appropriate roadbed. The article stated that the correct roadbed was a series of catenary sections and the square's center of gravity remained at the same height as the device rolled along. It also supplied a generic catenary function to serve as a construction template for the roadbed and went on to describe how to construct a sophisticated version of the device. Unfortunately, the author offered no further discussion or references regarding either the physics of the device or how he came to know that the correct roadbed was indeed a catenary. Believing that the device would probably yield some interesting physics and might be a good addition to my collection of demonstrations, I proceeded to investigate the device.

II. KINEMATICS

The kinematic problem of the square wheel is to determine a surface (curve) that allows the square to roll while maintaining a state of neutral equilibrium, as shown in Fig. 1. The prerequisite of neutral equilibrium means that the center of mass of the square must move along a horizontal trajectory and remain directly above the contact point. We will impose the restriction that rolling occurs without any slippage. These conditions suggest the geometry depicted in Fig. 2.

The square, assumed to be homogeneous, is described by sides of length $2a$ and center of mass at point A. The isosceles right triangle ABE represents one octant of the square. Set the zero of gravitational energy at the origin. Assume an initial configuration with one corner of the square located at the origin (point O) and the contiguous diagonal concurrent with the vertical (y) axis. In this configuration, the gravitational potential energy is proportional to the semidiagonal AB:$\{AB=R=\sqrt{2}a\}$. The angular displacement of the square is determined by the angle α. Point C is located on the perimeter of the square directly below A, and point D is the projection of the center of mass on the horizontal (x) axis. Pure rolling is ensured by the requirement that line segment BC=arc length OC.

Point C will generate the required curve when the gravitational potential energy of the square remains constant, that is $AC+CD=AB$. Let point C be described by its Cartesian coordinates (x,y). The critical kinematic relation becomes