Eternal Inflation and the Initial Singularity

Arvind Borde* and Alexander Vilenkin†
Institute of Cosmology, Department of Physics and Astronomy, Tufts University, Medford, Massachusetts 02155
(Received 28 October 1993)

It is shown that a physically reasonable spacetime that is eternally inflating to the future must possess an initial singularity.

PACS numbers: 98.80.Cq, 04.20.Dw

Inflationary models of cosmology [1] yield a rather different picture of the full present Universe (as opposed to the part of it that we can directly observe) than that given by the standard big bang cosmology. In current inflationary models the Universe consists of a number of isolated thermalized regions embedded in a still-inflating background [2]. The thermalized regions grow with time, but the inflating domains that separate them expand even faster, so that the Universe is never completely thermalized. It can be shown [3], that the boundary of a thermalized region is spacelike (and it lies to the past of points within the region), so that it is not possible to send a signal from the interior of a thermalized region to the "inflationary background" in which it is embedded.

Theoretical work supported by computer simulations suggests that this broad picture continues to describe the Universe as it evolves into the future [4]. Previously created regions expand and new ones come into existence, but the Universe does not fill up entirely with thermalized regions. In other words, inflation is *eternal to the future.*

A model in which the inflationary phase has no end and continually produces new islands of thermalization naturally leads to this question: Can this model also be extended to the infinite past, avoiding in this way the problem of the initial singularity? The Universe would then be in a steady state of eternal inflation without a beginning.

The purpose of this paper is to show that this is in fact not possible in future-eternal inflationary spacetimes as long as they obey some reasonable physical conditions: such models must necessarily possess initial singularities.

A partial answer along these lines was previously given by Vilenkin [5] who showed the necessity of a beginning in a two-dimensional spacetime and gave a plausibility argument for four dimensions. The broad question was also previously addressed by Borde [6] who sketched a general proof using the Penrose-Hawking-Geroch global techniques. The proof given below will partly follow the sketch outlined in that paper.

Statement of the result.—A spacetime cannot be past null geodesically complete if it satisfies the following conditions: (A) It is past causally simple. (B) It is open. (C) Einstein's equation holds, with a source that obeys the weak energy condition (i.e., the matter energy density is non-negative). (D) There is at least one point p such that for some point q to the future of p the volume of the difference of the pasts of q and p is finite.

Observe that geodesic incompleteness is being taken as a signal that there is a singularity. (A geodesic is incomplete if it cannot be continued to arbitrarily large values of its affine parameter.) This is the conventional approach in singularity theorems.

It is worth noting that none of the standard singularity theorems exactly fits the situation in which we are interested: some of the theorems assume the strong energy condition, known to be violated in inflationary scenarios, and others place much stronger restrictions on the global causal structure of spacetime than we do here [through assumption (A)].

More significantly, assumption (D) is entirely new—as we shall see below, it captures a characteristic aspect of future-eternal inflationary spacetimes.

Analysis of the assumptions.—Before we discuss the assumptions in detail, here is a summary: Assumptions (A) and (B) are made solely for mathematical convenience. The ultimate goal is to relax them [especially assumption (B)]. Assumption (C) holds in standard inflationary spacetimes, and is physically quite reasonable. Assumption (D) is necessary for inflation to be future-eternal.

In our detailed discussion of the assumptions, and in the proof given below, we will need to use some of the standard causal functions of the "global techniques" approach to general relativity. This is what we will need: A curve is called *causal* if it is everywhere timelike or null (i.e., lightlike). Let p be a point in spacetime. The *causal* and *chronological* pasts of p, denoted, respectively, by $J^-(p)$ and $I^-(p)$, are defined as follows: $J^-(p) = \{q$: there is a future-directed causal curve from q to $p\}$, and $I^-(p) = \{q$: there is a future-directed timelike curve from q to $p\}$.

The *past light cone* of p may then be defined as $E^-(p) = J^-(p) - I^-(p)$. It may be shown [7] that the boundaries of the two kinds of pasts of p are the same; i.e., $\dot{J}^-(p) = \dot{I}^-(p)$. Further, it may be shown that $E^-(p) \subset \dot{I}^-(p)$. In general, however, $E^-(p) \neq \dot{I}^-(p)$; i.e., the past light cone of p (as we have defined it here) is a subset of the boundary of the past of p, but is not necessarily the full boundary of this past. This is illustrated in Fig. 1.

With this background in hand, we may now discuss the assumptions in greater detail.

0031-9007/94/72(21)/3305(4)$06.00

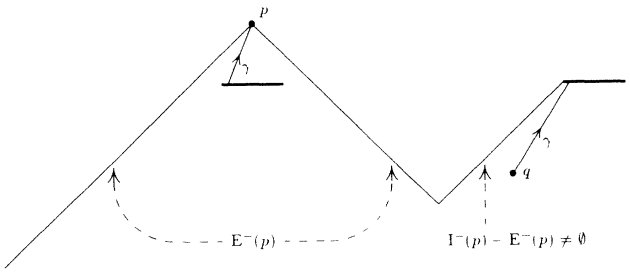

FIG. 1. An example of a spacetime that is not causally simple. The two thick horizontal lines are identified, allowing the point q to send a signal to p as shown (along the curve γ). The boundary of the past of p then consists of more than just the past light cone of p; i.e., $I^-(p)-E^-(p)$ is not empty.

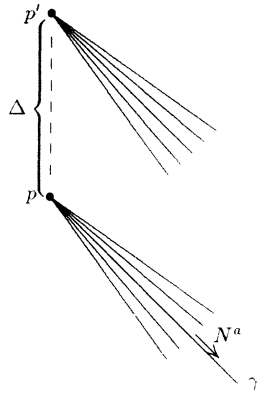

FIG. 2. The volume of interest.

Assumption (A): A spacetime is past causally simple if $E^-(p)=I^-(p)\neq\varnothing$ for all points p. I.e., we exclude for now scenarios such as the one in Fig. 1.

Assumption (B): A universe is open if it contains no compact, achronal, edgeless hypersurfaces. (A set is achronal if no two points in it can be connected by a timelike curve.)

Assumption (C): Einstein's equation is $R_{ab}-\frac{1}{2}g_{ab}R=kT_{ab}$ (where R_{ab} is the Ricci tensor associated with the metric g_{ab}, R the scalar curvature, T_{ab} the matter energy-momentum tensor, and k a positive constant in our conventions). An observer with four-velocity V^a will see a matter energy density of $T_{ab}V^aV^b$. The weak energy condition is the requirement that $T_{ab}V^aV^b\geq 0$ for all timelike vectors V^a.

This is a reasonable restriction on spacetime. (In fact, a much weaker integral condition will do just as well for our purposes [8].) It follows by continuity from the weak energy condition that $T_{ab}N^aN^b\geq 0$ for all null vectors N^a and from Einstein's equation that $R_{ab}N^aN^b\geq 0$. It is this final form, sometimes called the *null convergence condition*, that we will actually use. (Thus our results will remain true even in other theories of gravity, as long as the null convergence condition continues to hold.)

Assumption (D): This condition was formulated [5] as a necessary condition for inflation to be future-eternal. Consider a point p that lies in the inflating region; for inflation to be future-eternal there must be a nonzero probability for there to be a point q, at a given timelike geodesic distance δ to the future of p, that also lies in the inflating region. Since the boundary of a thermalized region is spacelike (and it lies to the past of points within the region), it is not possible to send a signal from the interior of a thermalized region to an inflating region. Thus, if a point r lies in a thermalized region, then all points to its future [i.e., all points in $I^+(r)$] are also thermalized. This means that there should be no thermalization events in the region $I^-(q)-I^-(p)$. Now, the formation of thermalized regions is a stochastic process, and it is reasonable to assume that there is a zero probability for no such regions to form in an infinite spacetime volume. This leads to assumption (D) [9]. For a more detailed discussion of assumption (D) see Ref. [5].

Proof.—Suppose that a spacetime that obeys assumptions (A)-(D) is past null geodesically complete. We show in two steps that a contradiction ensues.

(1) A point p that satisfies assumption D has a finite past light cone. By a "finite past light cone" is meant a light cone $E^-(p)$ such that every past-directed null geodesic that initially lies in the cone leaves it a finite affine parameter distance to the past of p.

Suppose, to the contrary, that a null geodesic γ lies in $E^-(p)$ an infinite affine parameter distance to the past of p. Let v be an affine parameter on γ chosen to increase to the past and to have the value 0 at p. Consider a small "conical" pencil of null geodesics in $E^-(p)$ around γ. Vary this pencil an infinitesimal distance Δ in the future time direction so that the vertex is now at the point p'. This sets up a congruence of null geodesics; let N^a be the tangent vector field to these geodesics. (See Fig. 2.)

How do the null geodesics in this congruence move away from or towards γ? If one considers a "deviation vector" Z^a that connects points on γ to points on nearby geodesics [7], a straightforward calculation yields the result that the only physically relevant variations in Z^a come from its components in the spacelike two-space in $E^-(p)$ that is orthogonal to N^a [10].

This means that the volume of the spacetime region occupied by the geodesic congruence may be expressed as

$$\delta V=\Delta\int_0^\infty \mathcal{A}(v)dv\,,$$

where $\mathcal{A}(v)$ is the area of the spacelike cross section of the light cone orthogonal to N^a. The region whose volume we are calculating is a subset of $\overline{I^-(q)-I^-(p)}$ (i.e., the closure of the difference of the pasts of q and p) and it must thus have a finite volume. In order for this to happen, $\mathcal{A}$ must decrease somewhere along γ.

The propagation equation for $\mathcal{A}$ is [7]

$$\dot{\mathcal{A}}=\theta\mathcal{A}\,,$$

where a dot represents a derivative with respect to v and $\theta = D_a N^a$ is the divergence of the congruence. If $\mathcal{A}$ decreases it follows that θ must become negative. But the propagation equation for θ may be written as [7]

$$\dot{\theta} \leq -\tfrac{1}{2}\theta^2 - R_{ab}N^aN^b \leq -\tfrac{1}{2}\theta^2$$

[where we have used assumption (C) in the last step]. If $\theta < 0$ somewhere, it follows that $\theta \to -\infty$ within a finite affine parameter distance.

The divergence of θ to $-\infty$ is a signal that the null geodesics from p have refocused. It is a standard result in global general relativity [7] that points on such null geodesics beyond the focal point enter the interior of the past light cone [i.e., enter $I^-(p)$] and no longer lie in $E^-(p)$. Thus the null cone $E^-(p)$ must be finite in the sense defined above.

(2) The result of step (1) contradicts assumption (B). From causal simplicity it follows that $E^-(p)$ [being equal to the full boundary of the past of p, $I^-(p)$] is an edgeless surface. It is also achronal. [If two points on it can be connected by a timelike curve, then the pastmost of the two points will lie inside $I^-(p)$, not on its boundary [7].] And step (1) implies that $E^-(p)$ is compact. These three statements taken together contradict assumption (B).

Discussion.—The conclusion to be drawn from this argument is that inflation does not seem to avoid the problem of the initial singularity (although it does move it back into an indefinite past). In fact, our analysis of assumption (D) suggests that almost all points in the inflating region will have a singularity somewhere in their pasts. As with most singularity theorems, our analysis tells us nothing about the nature of the singularity or about its precise location. In particular, we cannot tell whether or not the singularity occurs on a spacelike surface, like the big bang singularity of the standard Robertson-Walker cosmology. However, the fact that inflationary spacetimes are past incomplete forces one to address the question of what, if anything, came before. The most promising way to deal with this problem is probably to treat the Universe quantum mechanically and describe it by a wave function rather than by a classical spacetime.

The theorem that we have proved in this paper is based on several assumptions which it would be desirable to further justify or relax. The principal relaxation that is necessary is in assumption (B); i.e., closed universes must also be accommodated. It may appear at first sight that this will be difficult to achieve since assumption (B) entered into the proof above at a crucial place. However, all that we really need is to exclude situations where null geodesics recross after running around the whole Universe (as they do in the static Einstein model). If the reconvergence of the null cone discussed above occurs on a scale smaller than the cosmological one, then essentially the same argument goes through. This approach to applying open universe singularity theorems to closed universes has been outlined previously by Penrose [11] and it will be discussed in detail separately as will the relaxation of assumption (A).

Assumption (D), which plays a central role in our argument, requires further justification. It rests on the assumption that the probability of finding no thermalized regions in an infinite spacetime volume vanishes. The assumption is plausible, and in the original inflationary scenario [12,13], it is known to be true. It would be interesting, however, to determine the exact conditions of validity of assumption (D) and to investigate the possibility of relaxing it.

One of the authors (A.B.) thanks the Institute of Cosmology at Tufts University for its warm hospitality over the period when this work was done. The other author (A.V.) acknowledges partial support from the National Science Foundation.

*Permanent addresses: Long Island University, Southampton, NY 11968, and High Energy Theory Group, Brookhaven National Laboratory, Upton, NY 11973. Electronic mail: borde@bnlcl6.bnl.gov

†Electronic mail: avilenki@pearl.tufts.edu

[1] For reviews see, for example, A.D. Linde, *Particle Physics and Inflationary Cosmology* (Harwood Academic, Chur Switzerland, 1990); E. W. Kolb and M. S. Turner, *The Early Universe* (Addison-Wesley, New York, 1990).

[2] The inflationary expansion is driven by the potential energy of a scalar field φ, while the field slowly "rolls down" its potential $V(\varphi)$. When φ reaches the minimum of the potential this vacuum energy thermalizes, and inflation is followed by the usual radiation-dominated expansion. The evolution of the field φ is influenced by quantum fluctuations, and as a result thermalization does not occur simultaneously in different parts of the Universe. Fluctuations in the thermalization time give rise to small density fluctuations on observable scales, but result in large deviations from homogeneity and isotropy on much larger scales.

[3] The spacelike character of the boundaries of thermalized regions is a general feature of slow rollover inflation. More generally, in the slow-rollover regime the surfaces of constant field φ are spacelike. To show this, we note that during the slow rollover the field φ varies at the rate $\dot{\varphi} \gg H^2$, where H is the local expansion rate of the Universe [1]. Quantum fluctuations cause φ to vary by $\sim H$ on the horizon scale H^{-1}; the resulting spatial gradients are of the order $|\partial_i\varphi\partial^i\varphi| \sim H^4$. Hence, $\partial_\mu\varphi\partial^\mu\varphi > 0$ and the surfaces of constant φ are spacelike [our metric has signature $(+,-,-,-)$]. The situation here is different from that in the old inflationary scenario [12,13] where the bubble walls expand at speeds approaching the speed of light and their worldsheets are timelike surfaces. Note that although the boundaries of thermalized regions are spacelike surfaces, they are not Cauchy surfaces for the entire spacetime, and therefore the Universe can con-

tain a large number of nonoverlapping thermalized regions.

[4] P. J. Steinhardt, in *The Very Early Universe*, edited by G. W. Gibbons, S. W. Hawking, and S. T. C. Siklos (Cambridge University Press, Cambridge, England, 1983); A. Vilenkin, Phys. Rev. D **27**, 2848 (1983); A. A. Starobinsky, in *Field Theory, Quantum Gravity and Strings*, Proceedings of the Seminar Series, Meudon and Paris, France, 1984-1985, edited by M. J. de Vega and N. Sanchez, Lecture Notes in Physics Vol. 246 (Springer-Verlag, New York, 1986); A. D. Linde, Phys. Lett. B **175**, 395 (1986); M. Aryal and A. Vilenkin, Phys. Lett. B **199**, 351 (1987); A. S. Goncharov, A. D. Linde, and V. F. Mukhanov, Int. J. Mod. Phys. A **2**, 561 (1987); K. Nakao, Y. Nambu, and M. Sasaki, Prog. Theor. Phys. **80**, 1041 (1988); A. Linde, D. Linde, and A. Mezhlumian, Phys. Rev. D **49**, 1783 (1994).

[5] A. Vilenkin, Phys. Rev. D **46**, 2355 (1992).

[6] A. Borde, Classical Quantum Gravity **4**, 343 (1987).

[7] For the detailed proofs of such standard global results, see, for example, S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-Time* (Cambridge University Press, Cambridge, England, 1973).

[8] F. J. Tipler, J. Diff. Eq. **30**, 165 (1978); also see Ref. [6] for an extension of these results.

[9] Note that we are not assuming that there is a nonzero probability for a point p to have a finite-volume past [i.e., a finite volume for $I^-(p)$]. Such an assumption would mean that an arbitrarily chosen point has a finite probability to be in the inflating region. It has been shown, however, that at least in some models the inflating region is a fractal of dimension smaller than 4, and therefore occupies a vanishing fraction of the total volume (see Ref. [4]).

[10] This may be seen by introducing a pseudo-orthonormal basis $\{N^a, L^a, X^a_1, X^a_2\}$ where L^a is a null vector such that $N^a L_a = 1$ [in a convention where the metric has signature $(+,-,-,-)$] and X^a_1 and X^a_2 are unit spacelike vectors orthogonal to N^a and L^a. The deviation vector Z^a may be written as $Z^a = nN^a + lL^a + x_1 X^a_1 + x_2 X^a_2$. Now, it is possible to choose the affine parametrization so that $n = 0$. Further, $l = Z^a N_a$ and the derivative of l vanishes in the direction of N^a [i.e., $N^b D_b (Z^a N_a) = 0$].

[11] R. Penrose, in *Battelle Rencontres*, edited by C. M. DeWitt and J. A. Wheeler (Benjamin, New York, 1968).

[12] A. H. Guth, Phys. Rev. D **23**, 347 (1981).

[13] K. Sato, Mon. Not. R. Astron. Soc. **195**, 467 (1981).