THE MATHEMATICS OF THE BIG BANG

DANIEL J OLDER

Abstract. This paper aims to prove the Hawking singularity theorem, a theorem of Lorentzian geometry that has a very crucial implication in physics, namely proving the existence of the big bang singularity. Assuming certain conditions satisfied by our universe such as the predictability of the past and future, the limited speed of matter and energy, and the expansion of space translated into the language of Lorentz manifolds, an inevitable consequence is that any path a particle has traveled to get to this current moment in time cannot be longer than a fixed upper bound which can be interpreted as an upper bound on the age of the universe.

Contents

1. Preliminary Structures

We must first understand the structure of spacetime and a good mathematical model for spacetime turns out to be a Lorentz manifold. A Lorentz manifold is simply a smooth manifold with an added geometric structure that models spacetime quite well. We assume that the reader is familiar with the definitions of smooth manifold and the tangent space at a point on a smooth manifold. We will bypass the formal definitions of these structures and assume an understanding of them throughout the rest of this paper. Suffice it to say that a smooth n-manifold M is an object that locally looks like $\mathbf{R}^n$ and the tangent space at a point on M is the space

Date: August 26, 2013.

in which all the tangent vectors lie. These tangent vectors can represent velocity vectors, electromagnetic field vectors, or other physically significant vectors.

While we said we would not go into the formal definitions of tangent vectors, we would still like to clarify two different ways they are defined since these two definitions will be used interchangeably throughout this paper. One way is to look at the set of curves in M passing through p and to define an equivalence class over the curves such that all curves passing through p which are tangent to one another are equivalent. This allows us to look at velocities to curves as tangent vectors. The second way to view tangent vectors at p is as derivations, that is as directional derivatives acting on C^∞ functions defined on a neighborhood of p. We then say that the tangent space at a point $p \in M$, denoted T_pM, is just the vector space in which all the tangent vectors to M at p lie. We define the tangent bundle TM as the set of all tangent vectors on the manifold. The tangent bundle TM has a natural smooth structure inherited from M.

Another important notion is that of a smooth vector field. A smooth vector field X is an assignment of a single vector in the tangent space T_pM for each $p \in M$, such that it is a smooth section $M \mapsto TM$. The vector space of smooth vector fields on M we denote $\chi(M)$. If we have a neighborhood $V \subset M$ and a set $E_i \in \chi(V)$ such that the E_i's form a basis of T_pM for each $p \in V$, we call $\{E_i\}$ a frame field.

These are all the essential features of our smooth manifold structure for spacetime, but we also have a geometric structure associated with spacetime that we need to discuss. What endows a vector space with geometric properties of lengths and angles is the inner product. Thus we assign an inner product $\langle \cdot, \cdot \rangle$ to each tangent space which varies smoothly over the manifold in the sense that given $X, Y \in \chi(M)$, $\langle X, Y \rangle$ is a smooth function on M. We call $\langle \cdot, \cdot \rangle$ the metric on our manifold.

At every tangent space T_pM, if the largest subspace on which the metric is negative definite has dimension m, then the index of the metric is m. For our purposes, we will only care about metrics of index one. A smooth manifold equipped with a index one metric is called a Lorentz manifold. We will call a vector space with an index one inner product a Lorentz vector space. For a Lorentz vector space, the largest negative definite subspace is of dimension one and represents the temporal dimension while the other dimensions represent spacial dimensions. From now on, M will always refer to a Lorentz manifold and $\langle \cdot, \cdot \rangle$ will always refer to an index one metric.

We now must move on to the properties of Lorentz manifolds that will most concern us in regards to the Hawking singularity theorem. In proving the theorem, we make three assumptions about the universe: it's future and past are predictable based on the present, matter and energy cannot travel faster than the speed of light, and the universe is expanding everywhere at the present moment. The next section on causality will extract the ideas of future and past and we will translate the first assumption into the language of Lorentz manifolds. The following two sections

will be concerned with curves on our manifold that represent paths particles take through spacetime and what properties such curves hold when maximizing length. The penultimate section introduces a notion of spacial expansion and elaborates on what the second and third assumptions mean for a Lorentz manifold. Finally, in the last section we prove the Hawking singularity theorem by showing that no particle's past is longer than a fixed constant provided that the universe satisfies the three assumptions we have mentioned.

2. Causality in Lorentz Manifolds

The goal in this section is to build up what space, time, future, and past mean for Lorentz manifolds. At the end of this section we will introduce the notion of a globally hyperbolic manifold and explain its relevance to one of the three assumptions we make about our universe.

On a given tangent space T_pM, there are three types of vectors: spacelike, timelike, and null vectors. The length of a vector we denote $|X| = |\langle X, X\rangle|^{\frac{1}{2}}$.

Definition 2.1. A vector $X \neq 0$ is:

$$\text{Spacelike if } \langle X, X\rangle > 0.$$
$$\text{Timelike if } \langle X, X\rangle < 0.$$
$$\text{Null if } \langle X, X\rangle = 0.$$

If $X = 0$, then X is considered spacelike.

Following from this definition, a curve $\alpha : [a, b] \to M$ is called spacelike, timelike, or null if all of its tangent vectors are spacelike, timelike, or null, respectively. A spacelike curve is a path you could trace out between two points at an instant in time. A timelike curve is simply a path that a material particle could take in spacetime. A null curve is a path that light would follow through spacetime. A curve which has either timelike or null tangent vectors at each point is called a causal curve.

Furthermore, a curve $\alpha : [a, b] \to M$ is called extendible if it has an endpoint $p \in M$ for which with any neighborhood V about p, there exists a t_0 such that for all $t > t_0$, $\alpha(t) \in V$. A curve that is not extendible is called inextendible.

Next, we must understand how to extract notions of future and past from our Lorentz manifold. First, we do so on a tangent space. Here $\mathfrak{T}$ will denote the set of timelike vectors in a Lorentz vector space V.

Definition 2.2. Let V be our Lorentz vector space and let $u \in \mathfrak{T}$. Then we can define two timecones. The first one $C(u)$ is defined as

$$C(u) = \{v \in \mathfrak{T} : \langle u, v\rangle < 0\}.$$

The second one, denoted $C(-u)$, is defined as

$$C(-u) = \{v \in \mathfrak{T} : \langle u, v\rangle > 0\}.$$

One corollary to these definitions is that the two time cones are disjoint. Furthermore, if two timelike vectors are in the same timecone, then their timecones are equal. Also, it can be shown that $\mathfrak{T} = C(u) \cup C(-u)$ for any timelike vector u. Thus $\mathfrak{T}$ is uniquely decomposed into two disjoint subsets of V (see [4, p. 143]). One of these subsets we can call the future timecone of the origin and the other we can call the past timecone of the origin. The idea behind these definitions is that if we have a material particle at the origin, then at an infinitesimal increment backwards in time the particle would be found in the past timecone and analogously, at an infinitesimal time forward in time, the particle would be somewhere in the future timecone.

If there exists a timelike vector field $X \in \chi(M)$, then we say that M is time orientable. If M is time orientable, then we can give a notion of future and past to our whole manifold. We define $C(X)$ to be the set of future timelike vectors and $C(-X)$ to be the set of past timelike vectors. Intuitively, future or past timelike curves are curves with solely future or past timelike tangent vectors, respectively. For the rest of this paper, M will always be time orientable.

If we take a snap shot of our spacetime at a given moment in time, then we have the whole universe in an instant. We call such a slice a Cauchy surface and define it as follows:

Definition 2.3. A Cauchy surface A is a hypersurface $A \subset M$ which is met exactly once by every inextendible causal curve.

We have reason to believe that our universe admits such a Cauchy surface A. Any Lorentz manifold that admits a Cauchy surface is called a globally hyperbolic manifold so our universe is likely globally hyperbolic. We next need a global notion of future and past on our manifold. This is accomplished by introducing Cauchy developments.

Definition 2.4. The future Cauchy development of the Cauchy surface A is

$$D^+(A) = \{p \in M : \text{every past inxtendible causal curve through p meets A}\}.$$

Similarly, we define the past Cauchy development as,

$$D^-(A) = \{p \in M : \text{every future inxtendible causal curve through p meets A}\}.$$

Intuitively, $D^+(A)$ is the set of all points in the future of A and $D^-(A)$ is the set of all points in the past of A. Given a globally hyperbolic manifold M and a Cauchy surface $A \subset M$, it is not hard to show that $D(A) = D^+(A) \cup D^-(A) = M$ (see [3, p. 201]). This implies that any particle at any point in M must have at some point

in its future or past been in A, which is what we mean when we say that the past and future are predictable. Thus the assertion that spacetime is globally hyperbolic is equivalent to the assertion that all events in spacetime are predictable. This is an assumption we need to make about our universe to prove the existence of the big bang.

3. Geometry on Lorentz Manifolds

In this section, we will be concerned with geometric notions on a Lorentz manifold such as covariant derivative, arclength, and curvature as well as other tools derived from these three basic ones. From this section, we will discover when curves locally maximize arclength which is crucial to the Hawking singularity theorem.

Perhaps the most important tool after the metric for developing geometry on Lorentz manifolds is the Levi-Civita connection. Essentially, it gives us the notion of covariant derivative (the rate of change of one vector field along the integral curve of another). With it, we build up essentially all the later tools and concepts of Lorentzian geometry.

Before defining it though, we need to know the definition of a Lie bracket. If $X, Y \in \chi(M)$, then $[X, Y] = XY - YX$ is the Lie bracket of X and Y. If viewing X and Y as directional derivatives, then it makes sense that this gives us another vector field as a result, that is $[X, Y] \in \chi(M)$.

Definition 3.1. The Levi-Civita connection is the unique mapping $\nabla : \chi(M) \times \chi(M) \to \chi(M)$, which we denote by $\nabla(X, Y) = \nabla_X Y$, satisfying the following four properties. The first two are:

$$\nabla_{fX_1+gX_2} Y = g\nabla_{X_1} Y + f\nabla_{X_2} Y. \tag{1}$$

$$\nabla_X(fY_1 + gY_2) = (Xf)Y_1 + (Xg)Y_2 + f\nabla_X Y_1 + g\nabla_X Y_2. \tag{2}$$

The last two are:

$$X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle. \tag{3}$$

$$\nabla_X Y - \nabla_Y X = [X, Y]. \tag{4}$$

Remark 3.2. The first and second properties define an affine connection. Affine connections simply give us some notion of derivatives of vector fields along curves. The third property is called compatibility with the metric and the fourth property is a property called torsion-free. $\nabla_X Y$ measures how Y is changing along an integral curve of X. For a given metric on the manifold, the Levi-Civita connection is the unique connection which satisfies properties (3) and (4). The existence and uniqueness of the Levi-Civita connection for a Lorentz manifold is called the fundamental theorem of Lorentzian geometry but we will not prove it in this paper (see [2, p. 1-2]). For the rest of this paper, ∇ will refer to the LC (Levi-Civita) connection.

Using the LC connection, we can now give a notion of parallel transport along a curve. To parallel transport two vectors along a curve means to move them smoothly in such a way as to not change their length or the angle between them (see [3, p. 35]). This creates an isometric isomorphism between pair of each tangent spaces along a curve once we pick a curve from one tangent space's origin to the other's origin. Thus, the LC connection gives us a way of relating the tangent spaces of two distant points on our manifold.

Definition 3.3. Given a curve $\alpha : [a, b] \to M$, and a vector $V_0 \in T_{\alpha(a)}M$, the parallel transport of V_0 along α at $\alpha(t)$ is the unique vector field $V(t) \in T_{\alpha(t)}M$ satisfying $\nabla_{\alpha'} V(t) \equiv 0$ and $V(0) = V_0$.

Remark 3.4. The existence and uniqueness of the parallel transport are given by the existence and uniqueness of solutions to ordinary first order differential equations.

Another key geometric idea that we will want to explore is the distance between two points. Since the metric has index one in a Lorentz manifold, we cannot make a metric space out of the manifold and have an exact notion of distance between two arbitrary points. However, all that we care about in this paper is the length of timelike curves, which is the passage of time a particle would feel that was traveling along the curve.

Definition 3.5. Given a timelike curve $\alpha : [a, b] \to M$, the arclength of the curve, denoted $L(\alpha)$, is given by the formula:

$$L(\alpha) = \int_a^b (-\langle \alpha'(t), \alpha'(t) \rangle)^{\frac{1}{2}} \, dt. \tag{5}$$

An immediate question which arises is what properties might timelike curves have that locally extremize arclength. To find out, let $Q = (-\epsilon, \epsilon) \times [a, b]$ and let us take a one parameter family of curves $\alpha : Q \to M$, where each curve is written as $\alpha_s = \alpha(s, \cdot) : [a, b] \to M$ for fixed $s \in (-\epsilon, \epsilon)$. We will call the tangent vector fields on Q corresponding to the directional derivatives along the first and second variables S and T and naturally identify them with their images under the differential of α (see [2, p. 4]). S and T evidently commute so we know that $[S, T] = 0$, which means $\nabla_S T = \nabla_T S$ by the LC Connection being torsion-free. Furthermore, $l_s = L(\alpha_s)$ denotes the length of $\alpha(s, [a, b])$. By parameterizing the curve by arclength, we can

thus make $|T| \equiv l_s$ for each s.

$$\begin{aligned}
\frac{\partial}{\partial s} L(\alpha_s)\Big|_{s=0} &= \int_a^b S(-\langle T, T\rangle)^{\frac{1}{2}}\, dt. \\
&= \int_a^b -\frac{1}{|T|}\langle \nabla_S T, T\rangle\, dt. \\
&= \frac{1}{l_0}\int_a^b \langle \nabla_T S, T\rangle\, dt. \\
&= \frac{1}{l_0}\int_a^b -T\langle S, T\rangle + \langle S, \nabla_T T\rangle\, dt. \\
&= \frac{1}{l_0}\Big(-\langle S, T\rangle\Big|_a^b + \int_a^b \langle S, \nabla_T T\rangle\, dt.\Big) \\
&= \frac{1}{l_0}\int_a^b \langle S, \nabla_T T\rangle\, dt. \qquad (6)
\end{aligned}$$

The above will be equal to zero for all possible variations S if and only if $\nabla_T T \equiv 0$ for the curve α_0. A timelike curve $\gamma : [a, b] \to M$ which satisfies $\nabla_{\gamma'}\gamma' \equiv 0$ is called a timelike geodesic curve. From the first variation, we can see that the timelike curves that extremize arclength are timelike geodesics. The existence and uniqueness of solutions to second order ordinary differential equations imply that a point p and tangent vector $T_0 \in T_pM$ uniquely determine a geodesic γ_{T_0} such that $p = \gamma_{T_0}(0)$ and $T_0 = \gamma'_{T_0}(0)$ (see [1, p. 65]).

An important consequence of geodesics extremizing arclength is the following proposition. It is critical to our result at the end of this section on finding out which curves maximize arclength.

Proposition 3.6. A timelike geodesic from a Cauchy surface A to a point $p \in D^-(A)$ that is the longest possible curve from A to p must necessarily be perpendicular to A.

Proof. Since the curve is maximal, the first derivative of its arclength for any variation must be zero. Now consider a variation of our geodesic for which the variation vector field S on the geodesic satisfies $S(a) \in T_\gamma(a)A$. Since all curves in the variation end at the same point $p = \gamma(b)$, we know $S(b) = 0$. Thus, we arrive at the conclusion that

$$\frac{\partial}{\partial s} L(\alpha_s)\Big|_{s=0} = \frac{1}{l_0}(-\langle S, T\rangle)\Big|_a^b = 0$$

This means $\langle S(a), \gamma'(a)\rangle = 0$ and so γ is perpendicular to all vectors in $T_\gamma(a)A$ and thus perpendicular to A itself. □

Now that we have geodesics, we can define a very useful and important local diffeomorphism between the tangent space and the manifold in a neighborhood of a point p, namely the exponential map.

Definition 3.7. Let $p \in M$. The exponential map at p, $exp_p : T_p(M) \to M$, is defined by $exp_p(T_0) = \gamma_{T_0}(1)$ where γ_{T_0} is defined as was discussed above.

Since geodesics parallel transport their tangent vectors, their tangent vectors don't change length over the whole curve. Thus, if γ is parameterized by arclength, then $L(\gamma_{T_0}) = \gamma'_{T_0}(0) = |T_0|$. This means exp_p maps vectors in the tangent space T_pM of length l to points on the manifold a distance l away from p along the corresponding geodesic.

Given this fact, we can define something called a normal neighborhood about p. A normal neighborhood V about p is a neighborhood for which exp_p is a local diffeomorphism from a neighborhood $U \subset T_pM$ of 0 to $V \subset M$.

From this definition comes the normal balls B_ϵ and normal spheres S_ϵ of radius ϵ about a point $p \in M$. These balls and spheres on the manifold are the images of ϵ balls and spheres in the tangent space T_pM under the exponential map exp_p.

Another important geometric concept that arises naturally out of the definition of the exponential map is that of a conjugate point. We say that p is conjugate to q if exp_p is singular at q. While this is a useful definition, there is another useful definition of a conjugate point, which we will need later to really know what a conjugate point signifies geometrically. However, in order to get to this other definition, we must define two other geometric objects first: curvature and Jacobi fields.

Definition 3.8. The curvature on a manifold is a map $R : \chi(M) \times \chi(M) \times \chi(M) \to \chi(M)$ defined as:

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z. \tag{7}$$

Curvature measures the failure of covariant derivatives to commute and it will be useful in most later definitions and proofs.

Now a very important class of vector fields in Lorentzian geometry are Jacobi fields. Jacobi fields arise from one-parameter variations of geodesics where all of the curves in the variation are themselves geodesics. From this definition, arises the defining equation for Jacobi fields.

Definition 3.9. A Jacobi field is a vector field along a geodesic $\gamma : [0,1] \to M$ satisfying the following equation:

$$\nabla_{\gamma'} \nabla_{\gamma'} J = -R(\gamma', J)\gamma'. \tag{8}$$

Remark 3.10. From the existence and uniqueness of solutions to second order partial differential equations, a Jacobi field is uniquely determined by its value and

derivative at the start of a geodesic $\gamma(a)$ or by its value at the endpoints $\gamma(a)$ and $\gamma(b)$ as long as the endpoints are not conjugate points (see [2, p. 12]).

Now we may arrive at our alternative definition of a conjugate point.

Proposition 3.11. Let $\gamma : [a, b] \to M$ be a geodesic. There exists a nontrivial Jacobi field J defined on γ such that $J(a) = J(b) = 0$ if and only if $\gamma(a)$ and $\gamma(b)$ are conjugate points of each other.

Proof. Suppose $p = \gamma(a)$ and $q = \gamma(b)$ are conjugate to each other. Let $p_s(t) = (V + sW)t$ be a one parameter family of rays from the origin of T_pM outwards. Then, by the definition of the exponential map, $\alpha_s(t) = exp_p(p_s(t))$ is a one parameter family of geodesics starting at p. Let $\gamma(t) = p_0(t)$. Then $\frac{\partial}{\partial s}\alpha_0(b) = 0$, since p and q are conjugate. Using our usual vector fields for our variation S and T, we will call the restriction of S to the geodesic curve α_0 J. Thus, we know that $J(b) = 0$. The vector field J on α_0 is necessarily a Jacobi field on α_0 due to the construction of the variation and clearly $J(a) = 0$, which means $J(a) = J(b) = 0$.

To prove the other direction, suppose J is a nonzero Jacobi field on γ for which $J(a) = J(b) = 0$. Then let $\alpha(s, t) = exp_p(T + sJ'(a))t$ be a one parameter family of geodesics for which the variation vector field is J at $s = 0$ and T at p is just $\gamma'(0)$. Then

$$d(exp_p)_{\gamma'(0)}J'(a) = J(b) = 0.$$

Thus exp_p is singular at q and we are done. □

We wish to show that timelike geodesics are the timelike curves that locally maximize arclength when there are no conjugate points but before doing so, we must understand how to take the second derivative of arclength. The index form is the quadratic form that measures the second derivative of arclength when given a two parameter variation of a geodesic curve. Since we will only care about the second derivative of arclength of a one parameter variation, we will derive the index form from such a variation.

Suppose $\alpha_s : [a, b] \to M$ is a one parameter variation of the timelike geodesic α_0 where $s \in (-\epsilon, \epsilon)$. Let us call T the tangent vector field and S the variation vector

field as usual. We parameterize it such that $|T| \equiv 1$. Then

$$\begin{aligned}
\frac{\partial^2}{\partial s^2} L(\alpha_s)\Big|_{s=0} &= \frac{\partial}{\partial s} \int_a^b -\langle \nabla_T S, T\rangle \, dt. \\
&= \int_a^b -\langle \nabla_S \nabla_T S, T\rangle - \langle \nabla_T S, \nabla_S T\rangle \, dt. \\
&= \int_a^b \langle R(T,S)S, T\rangle - \langle \nabla_T \nabla_S S, T\rangle - \langle \nabla_T S, \nabla_S T\rangle \, dt. \\
&= \int_a^b \langle R(T,S)S, T\rangle - T\langle \nabla_S S, T\rangle - \langle \nabla_T S, \nabla_T S\rangle \, dt. \\
&= \langle \nabla_S S, T\rangle\Big|_a^b + \int_a^b \langle R(S,T)T, S\rangle - \langle \nabla_T S, \nabla_T S\rangle \, dt. = I(S,S). \quad (9)
\end{aligned}$$

Thus, we have defined the index form $I(S,S)$ for a variation S of a timelike geodesic.

There are a few key properties of the index form that establish the minimization of timelike geodesics. We shall list them in the following proposition and the proofs can be found in [2, p. 18, 20-22].

Proposition 3.12. If $\gamma : [a,b] \to M$ is a geodesic and I is the index form on vector fields defined on γ, then the following three propositions are true:

(1) I(J, J) = 0 if and only if J is a Jacobi Field.
(2) If J is a Jacobi Field and S is an arbitrary variation vector field such that $J(a) = S(a) = 0$ and $J(b) = S(b)$, then $I(J,J) \geq I(S,S)$.
(3) If there exists a point $t_0 \in (a,b)$ such that $\gamma(a)$ is conjugate to $\gamma(t_0)$, then we can find a vector field N along the geodesics for which $I(N,N) > 0$.

Now we are ready to finally prove which timelike curves in fact maximize arclength.

Lemma 3.13. If a curve $\gamma : [a,b] \to M$ goes from $p \in A$ to $q \in D^-(A)$ and it locally maximizes arclength, then it is a geodesic from the Cauchy surface A to the point q orthogonal to A with no conjugate points in between p and q.

Proof. Since γ maximizes arclength, it must be a geodesic. Since γ is a geodesic, it must be orthogonal to A. Suppose there are no conjugate points from p to q. Then by combining properties (1) and (2) of the index form from above, given a variation vector field S on γ such that $S(0) = 0$, the unique Jacobi field which aligns with S at the endpoints will satisfy $I(S,S) \leq I(J,J) = 0$. Thus I is negative definite on vector fields of γ if there are no conjugate points along γ. Suppose there is a conjugate point between p and q. By property (3), we can find a vector field N along γ such that $I(N,N) > 0$ This means that if γ maximizes arclength then there must not be any conjugate points along it between its endpoints. □

4. The Time Separation Function

The aim of this section is to show that there does exist a curve of maximal arclength between any two timelike separated points. This is absolutely crucial to our proof of the Hawking singularity theorem.

We would first like to quantify of distance between two timelike separated points using the time separation function τ.

Definition 4.1. We define the time separation $\tau : M \times M \to [0, \infty]$ of two points to be $\tau(p, q) = sup\{L(c) : c$ is a future pointing causal curve from p to q $\}$. If $A, B \subset M$, then the time separation $\tau(A, B) = sup\{\tau(a, b) : a \in A$ and $b \in B\}$.

Thus τ measures the maximal arclength between timelike separated points. Our aim in what follows will be to show that this function is continuous.

Lemma 4.2. If M is a globally hyperbolic spacetime, then τ is continuous.

Proof. First we will show that τ is lower semicontinuous. We know $\tau : M \times M \to [0, \infty]$. If $\tau(p, q) = 0$, then of course it will be lower semi-continuous at this (p, q). Now suppose $0 < \tau(p, q) < \infty$. We want to show given $\epsilon > 0$, there exist a small enough neighborhood about (p, q) such that $\tau(p', q') > \tau(p, q) - \epsilon$ for all (p', q') in that neighborhood. Fix $\epsilon > 0$ and let $\alpha : [a, b] \to M$ be a timelike curve from p to q such that $L(\alpha) > \tau(p, q) - \frac{\epsilon}{3}$. Take normal neighborhoods U about p and V about q and let $p_1 \in \alpha([a, b]) \cap U$ and $q_1 \in \alpha([a, b]) \cap V$. We will call the section of α from q to q_1 α_{qq_1}, from q_1 to p_1 $\alpha_{q_1p_1}$, and from p_1 to p α_{p_1p}. Since the function $(p, q) \mapsto L(\gamma_{pq})$ mapping two points to the length of the geodesic connecting them is clearly continuous on a normal neighborhood of M, we can take any $p' \in U$ and $q' \in V$ knowing that $L(\gamma_{p'p_1}) > L(\gamma_{pp_1}) - \frac{\epsilon}{3}$ and $L(\gamma_{q_1q'}) > L(\gamma_{q_1q}) - \frac{\epsilon}{3}$. Let us consider the curve α' from p' to q' made up of the segments $\gamma_{p'p_1}$, $\alpha_{p_1q_1}$, and $\gamma_{q_1q'}$. Using the fact that geodesics locally maximize causal curves and our previous inequalities, we have that

$$\begin{aligned}
\tau(p', q') \geq L(\alpha') &= L(\gamma_{p'p_1}) + L(\alpha_{p_1q_1}) + L(\gamma_{q_1q'}) \\
&> L(\gamma_{pp_1}) + L(\alpha_{p_1q_1}) + L(\gamma_{q_1q}) - \frac{2\epsilon}{3} \\
&\geq L(\alpha_{pp_1}) + L(\alpha_{p_1q_1}) + L(\alpha_{q_1q}) - \frac{2\epsilon}{3} \\
&= L(\alpha) - \frac{2\epsilon}{3} > \tau(p, q) - \epsilon.
\end{aligned}$$

Therefore, $\tau(p', q') > \tau(p, q) - \epsilon$. If $\tau(p, q) = \infty$, then take a curve α from p to q of length A and with a similar procedure to what we did above, we can show that $\tau(p', q') > A$. Thus, τ is lower semicontinuous.

Now to prove that τ is fully continuous, we must show that on a globally hyperbolic spacetime M, τ is upper semi-continuous. Arguing by contradiction, suppose there is a $p, q \in M$ such that τ is not upper semi-continuous at (p, q). Let $\{p_n\}$ and $\{q_n\}$ be sequences such that $p_n \to p$ and $q_n \to q$ and also such that for $\epsilon > 0$, we have that $\tau(p_n, q_n) \geq \tau(p, q) + \epsilon$. Let can pick $p^- \in D^-(p)$ and $q^+ \in D^+(q)$ and without loss of generality, suppose that $\{p_n\} \subset D^+(p^-)$ and $\{q_n\} \subset D^-(q^+)$. If we pick causal curves α_n from p_n to q_n such that $\tau(p_n, q_n) \geq L(\alpha_n) > \tau(p_n, q_n) - \frac{1}{n}$, then it follows that $\{\alpha_n\} \subset D^+(p^-) \cap D^-(q^+)$. A consequence of $C(p^-, q^+)$ being compact which we will not prove here is that $D^+(p^-) \cap D^-(q^+)$ is itself compact (see [3, p. 207]). This implies that there exists a limit curve λ of the α_n's which goes from p to q and has length $L(\lambda) \geq \tau(p, q) + \epsilon$. This contradicts the fact that τ is a supremum and so we thus conclude that τ is indeed upper semi-continuous on M as well and hence, is continuous. □

Now we would like to show that the space of continuous curves from p to q in M with the Hausdorff metric, denoted $C(p, q)$, is compact.

Lemma 4.3. If M is globally hyperbolic, $A \subset M$ is a Cauchy surface, and $p_0 \in D^-(A)$, then $C(A, p)$ is compact.

Proof. Let $\{\lambda_n\} \subset C(A, p_0)$ be an infinite sequence of curves. Take a convex normal neighborhood of p_0 and then a normal sphere S_ϵ within this normal neighborhood. Since the exponential map takes the sphere of radius ϵ in the tangent space at p_0 to S_ϵ and continuous functions map compact sets to compact sets, it follows that S_ϵ is compact.

Considering all the points where the λ_n's intersect S_ϵ, we know that there is a limit point on S_ϵ so let us call that point p_ϵ. We can likewise find a limit point $p_{a\epsilon}$ for each sphere $S_{a\epsilon}$ where $a \in \mathbf{Q} \cap [0, 1]$. The curves λ_n effectvely converge to a point at each fractional distance of ϵ away from p_0 in our normal ball B_ϵ about p_0. By taking the closure of this set of points, we end up with a continuous curve from p_0 to p_ϵ which is the limit curve of $\{\lambda_n\}$ (see [5, p. 210]).

We can extend such a limit curve further. First, we take a normal neighborhood B_{ϵ_2} of p_ϵ. Then we take a subsequence $\left\{_{\lambda_{n_i}}\right\}$ of $\{\lambda_n\}$ converging to p_ϵ. Taking p_ϵ to be our new p_0 and our subsequence $\{\lambda_{n_i}\}$ to be our new infinite sequence of curves, we can continue to extend our limit curve λ. The result is to build a smooth limit λ of the $\{\lambda_n\}$ that can be continued as far into the past as possible and hence it becomes past inextendible. Thus, $\lambda \in C(A, p_0)$ and we have shown that an arbitrary infinite subset of $C(A, p_0)$ has a limit point in $C(A, p_0)$ which shows that $C(A, p_0)$ is compact. □

From these two lemma's it follows that if τ is acting on $C(A, p)$, then since continuous functions admit maximums on compact sets, there exists a curve from A to

p which has the longest arclength of any such curve. This fact is very important in the proof of the Hawking singularity theorem.

5. Curvature and the Expansion of Universe

In this penultimate section, we will introduce two important tools, the Ricci curvature and the future convergence. By setting conditions on these two geometric objects that correspond to conditions that we believe exist in our universe, we can derive a very interesting result that is absolutely essential for the Hawking singularity theorem.

We begin by defining Ricci curvature.

Definition 5.1. Let $X, Y, Z \in \chi(M)$. The Ricci curvature is defined as the trace of the linear map $Z \mapsto R(Z, X)Y$. In an orthonormal frame field $\{E_i\}$, the Ricci curvature $Ric : \chi(M) \times \chi(M) \to \mathbf{R}$ can be computed as

$$Ric(X, Y) = \sum_{i=1}^{n} \langle R(E_i, X)Y, E_i \rangle. \tag{10}$$

We will be assuming an important condition that we believe holds true for the Universe known as the weak energy condition. It says that $Ric(X, X) \geq 0$ for all X timelike. This condition likely holds true for our universe because Einstein's Field Equations assert that the Ricci Curvature Ric is proportional to the energy-momentum tensor $\mathbf{T}$. Having $\mathbf{T}(X, X) \geq 0$ for all X timelike is the condition that nothing can travel faster than the speed of light through spacetime, which is a physically plausible assumption.

The next object we will consider is the future convergence but first we must define two preliminary objects, the second fundamental form and the mean curvature vector field.

Given a submanifold $\bar{M}$ of a Lorentz manifold M, if we consider two vector fields $V, W \in \chi(\bar{M})$, we can break up the covariant derivative of W with respect to V into parallel and perpendicular components to $\bar{M}$. We call the parallel component the induced connection $\bar{\nabla}$ on $\bar{M}$, since it can clearly be shown to be the unique LC connection on the submanifold $\bar{M}$. We call the perpendicular component the second fundamental form, denoted $II(V, W)$. This can all be written succinctly by the formula

$$\nabla_V W = \bar{\nabla}_V W + II(V, W). \tag{11}$$

Essentially, if our submanifold is a Cauchy surface $A \subset M$, the second fundamental form measures how a normal vector to A fails to be parallel transported along the given direction with respect to the LC connection on M (see [5, p. 46]). With the

second fundamental form, we can then define the mean curvature vector field on a sub-manifold as follows:

Definition 5.2. Given a Cauchy surface A and the future unit normal vector field U on A, the mean curvature vector field H is defined as the trace of the linear map $U \mapsto \langle II(V,W), U\rangle U$. With an orthonormal frame $\{E_i\}$ on A, H can be computed as

$$H = \frac{1}{n-1}\sum_{i=1}^{n-1} II(E_i, E_i). \tag{12}$$

If we picture a one-parameter family of Cauchy surfaces $\{A_t\}$ dividing up our manifold M such that each A_t represents all of space at a moment in time t, then mean curvature essentially measures the derivative of spacial volume with respect to time as we move towards the future in M (see [3, p. 230]). It would take us too far afield to explicitly discuss how this definition can be understood in this way but we will use this as our intuition for mean curvature going forward.

The future convergence is simply the magnitude of the mean curvature vector field and thus measures at what rate the volume of space is expanding or contracting in at a given moment in time.

Definition 5.3. If U is the future unit normal to the Cauchy surface A, then the future convergence k is defined as

$$k(U) = \langle U, H\rangle. \tag{13}$$

We currently believe that our universe is expanding based on the acceleration away from us of galaxies in all directions. This condition on our universe is summed up succinctly by the formula $k(U) > 0$ where U is defined over a whole Cauchy surface A. In other words, volume at every point in space at the present moment in time is expanding.

Now we can finally move on to the main lemma of this section which is critical to proving the Hawking singularity theorem.

Lemma 5.4. If $\gamma : [a,b] \to M$ is a geodesic from $p \in A$ normal to A and the following two conditions are satisfied:

$$k = k(\gamma'(0)) > 0. \tag{14}$$

$$Ric(\gamma', \gamma') \geq 0. \tag{15}$$

then there exists a conjugate point $\gamma(r)$ to $\gamma(0)$ at $0 < r \leq \frac{1}{k}$ as long as γ is still defined on this interval.

Proof. Let $e_i \in T_pS$ be an orthonormal frame of spacelike vectors and E_i be it's parallel translation along γ. Let $f(t) = 1 - kt$ on $[0, \frac{1}{k}]$. Consider the frame field $F_i = fE_i$ of spacelike vectors along γ. Now looking at the index form of F_i, we have:

$$\begin{aligned} I(F_i, F_i) &= \langle -\nabla_{F_i} F_i, T\rangle \Big|_0^{\frac{1}{k}} + \int_0^{\frac{1}{k}} \langle R(F_i, T)T, F_i\rangle - \langle \nabla_T F_i, \nabla_T F_i\rangle \, dt \\ &= f(0)^2 \langle \nabla_{e_i} e_i, T\rangle + \int_0^{\frac{1}{k}} \langle R(F_i, T)T, F_i\rangle - f'(t)^2 \langle E_i, E_i\rangle \, dt \\ &= \langle II(e_i, e_i), T\rangle + \int_0^{\frac{1}{k}} \langle R(F_i, T)T, F_i\rangle \, dt - k. \end{aligned}$$

If we then sum over our frame vector fields F_i, we get

$$\begin{aligned} \sum_{i=1}^{n-1} I(F_i, F_i) &= \langle (n-1)H, T\rangle + \int_0^{\frac{1}{k}} f^2(t) Ric(T, T) \, dt - (n-1)k \\ &= \int_0^{\frac{1}{k}} f^2(t) Ric(T, T) \, dt \geq 0. \end{aligned}$$

This implies that $I(F_i, F_i) \geq 0$ for at least one F_i. This however means that γ is no longer maximizing and thus, there must be a conjugate point $\gamma(r)$ with $r \in (0, \frac{1}{k})$. $\square$

6. The Hawking Singularity Theorem

Finally, we have reached the pinnacle of the paper. The main idea behind the Hawking singularity theorem is that we take an arbitrary point in our past and consider the longest timelike curve going from that point to the present time. If we bound the length of such a curve from below by a particular bound relating to the rate of expansion of our Universe, we necessarily find that the timelike curve admits a conjugate point. This, however, negates it being maximal and leads to the conclusion that the curve along with all timelike curves reaching into the past must be bounded from above. This essentially means nothing reaches back into the past beyond a certain finite time and thus, the universe as we know it must have had a beginning!

Theorem 6.1. Let M be a globally hyperbolic spacetime and let $Ric(X, X) \geq 0$ for all timelike vectors X. If A is a Cauchy surface on which the future convergence at every point satisfies that $k \geq C > 0$ where C is a constant, then every past-directed timelike curve from A must not have a length greater than $\frac{1}{C}$.

Proof. Consider $q \in D^-(S)$. Since $C(A, q)$ is compact and τ is continuous, τ takes on a maximum value in $C(S, q)$. Let us call that maximum length curve $\gamma : [a, b] \to M$ where $p = \gamma(a) \in S$. Suppose that $L(\gamma) > \frac{1}{C}$. Since timelike curves that locally

maximize distance must by necessity be orthogonal geodesics with no conjugate points, it follows that γ must be a geodesic orthogonal to S with no conjugate points between p and q. But given the conditions on Ricci curvature and the convergence on A, it follows that γ must have a conjugate point between p and q. This is a contradiction. Thus, our initial assumption that we could find a timelike curve of length longer than $\frac{1}{C}$ was wrong. This means all past directed timelike curves starting at S have length less than or equal to $\frac{1}{C}$. □

7. Acknowledgements

I would like to thank my mentor Clark Butler for being a terrific mentor for this summer. He continually gave me good books and exercises that really helped me learn the subject matter. Even though Lorentzian geometry was a subject a bit outside either of our comfort zones, he stuck with me and helped me work through the material.

References

[1] do Carmo, Manfredo. Riemannian Geometry. Boston, Massachusetts, Berkhäuser Boston, 1993.
[2] Cheeger, Jeff and Ebin, David G. Comparison Theorems in Riemannian Geometry. Providence, Rhode Island, AMS Chelsea Publishing, 1975.
[3] Wald, Robert General Relativity Chicago, Illinois, University of Chicago Press, 1984.
[4] O'Neil, Barrett Semi-Riemannian Geometry with Applications to Relativity. New york, New York, Academic Press, Inc., 1983.
[5] Hawking, S.W. and Ellis, G.F.R. The Large Scale Structure of Space-time. London, United Kingdom, Cambridge University Press, 1973.