UNIVERSITY OF CALIFORNIA

Santa Barbara

The Big Bang Singularitry

A Thesis submitted in partial satisfaction of the

requirements for the degree Master of Arts

in Mathematics

by

Eric Ling

Committee in charge:

Professor Xianzhe Dai, Chair

Professor Rick Rugang Ye

Professor Guofang Wei

June 2015

The thesis of Eric Ling is approved.

Rick Rugang Ye

Guofang Wei

Xianzhe Dai, Committee Chair

May 2015

ACKNOWLEDGMENTS

I wish to express my sincere thanks to all members of the Department of Mathematics at UC Santa Barbara for their help and support. I am especially grateful to my advisor Xianzhe Dai for guiding me towards the mathematical topics in general relativity. I also thank my parents for their encouragement and support.

ABSTRACT

The Big Bang Singularity

by

Eric Ling

The big bang theory is a model of the universe which makes the striking prediction that the universe began a finite amount of time in the past at the so called "Big Bang singularity." We explore the physical and mathematical justification of this surprising result. After laying down the framework of the universe as a spacetime manifold, we combine physical observations with global symmetrical assumptions to deduce the FRW cosmological models which predict a big bang singularity. Next we prove a couple theorems due to Stephen Hawking which show that the big bang singularity exists even if one removes the global symmetrical assumptions. Lastly, we investigate the conditions one needs to impose on a spacetime if one wishes to avoid a singularity. The ideas and concepts used here to study spacetimes are similar to those used to study Riemannian manifolds, therefore we compare and contrast the two geometries throughout.

1 Introduction

We can describe events in our universe by four coordinates: three to describe where we are and one to describe when we are. Our knowledge of the universe is limited by what we can measure and observe near us, so we only have a local understanding of the universe. Here local could mean the observable universe, which is large, but nonetheless local. Thus we can describe the universe as a four-dimensional topological manifold M (see [4] for the relevant definitions). We can add extra structure to M based on our everyday experience. For example, the use of calculus in our everyday lives suggest that M should possess a smooth (or at least highly differentiable) structure. Likewise, our observations suggest that M satisfies the Hausdorff separation axiom. We also assume M is connected since we would have no knowledge of any disconnected component. The last and most important structure that M is equipped with is a Lorentzian metric g. This means that for every point $p \in M$, there is a basis $\{e_0, e_1, e_2, e_3\}$ in T_pM such that the components of g in this basis are $g_{ab} = g(e_a, e_b) = \text{diag}[-1, 1, 1, 1]$. The Lorentzian metric is very important but not intuitive to understand from our everyday experiences. Because of this, we dedicate this section to motivating it.

Galileo was the first to suggest that motion was relative. Imagine a person A standing still on the Earth and a person B moving in a horse carriage. Galileo would say, yes person B is moving but only relative to person A; it's equally valid to say person A is moving relative to person B. Moreover, if B's motion was constant, then B would be unable to determine if B was moving or not provided the carriage has no windows. A

and B have their own frame of reference, that of the Earth and the carriage, respectively. Galileo eliminated the idea that the Earth was a special reference frame. This was not all obvious at the time. Everyday experience would suggest that the Earth's reference frame was special because all objects in motion eventually stop moving. For example, a ball initially thrown will eventually come to rest with respect to Earth's reference frame. We now understand that this is due to frictional forces from the air and ground. Thus, Galileo established that motion is relative; there is no preferred reference frame.

In the 1800s a lot of experimental and theoretical research in physics went into to describing electric and magnetic phenomenon. The culmination of this work led to the pervasiveness of Maxwell's equations. It was soon discovered that these equations imply a three-dimensional wave equation. These waves came to be known as electromagnetic radiation and they coincidentally traveled at speed c, the speed of light. It was soon realized that light itself is electromagnetic radiation. There was a serious problem though. Maxwell's equations do not specify which reference frame we are to consider for the speed of light, and the fact that such a reference frame exists means that Galileo was wrong: there is a preferred reference frame - the one we use to calculate c. For example, is the speed c to be taken in the reference frame of the Earth or the reference frame of a train traveling on the Earth? It was well known at the time that the Earth revolves around the Sun so its reference frame is not inertial. The only special reference frame that seemed to be inertial was the Sun's frame (at the time scientists did not know the Sun revolves around the core of a Galaxy).

Einstein took another route. He believed that Galileo was right and that there are no preferred reference frames. This means that the speed c predicted by Maxwell's equa-

tions must be measured by any observer in any reference frame. What does this mean physically? Suppose observer A is standing on Earth and observer B is on a train moving at a speed v relative to A. At the moment B passes A, both B and A shine a flashlight in the direction of the train's motion. The photons from both B and A's flashlight will be traveling next to each other, neither passing the other. Before Einstein it was believed that B's photons would travel at a speed v *faster* than A's photons, as one would expect from everyday experience.

Let's suppose A labels his time coordinate by t and the distance in the direction of the train by x. If B labels his time coordinate by t' and the distance in the direction of the train by x'. We want to find a relationship between (t, x) and (t', x'). Before Einstein, the relationship was trivial

$$t' = t \quad \text{and} \quad x' = x + vt.$$

We have to find a new relationship that incorporates Einstein's belief that both A and B will measure the same speed of light. Suppose the relationship we seek is of the form

$$t' = \alpha x + \beta t \quad \text{and} \quad x' = \gamma x + \delta t,$$

where α, β, γ, and δ are to be determined. Suppose x is measuring the position of the train. Since the train is moving with a speed v relative to A, $x = vt$, and since the train does not move for B, $x' = 0$. Plugging these into our formula for x', we get $\gamma vt + \delta t = 0$, therefore $\delta = -\gamma v$. Now suppose x measures the position of A, then $x = 0$ but $x' = -vt$. Then

$$-v(0 + \beta t) = -vt' = x' = \gamma(x - vt) = \gamma(0 - vt).$$

Therefore $\beta = \gamma$. Now suppose a light pulse is sent out by A from the origin along the x axis at $t = 0$. Einstein believed that A measures the location of the light pulse as $x = ct$ and B measures the location of the light pulse as $x' = ct'$ (as oppose to $x' = ct' - vt'$). Then

$$\gamma(ct - vt) = \gamma(x - ct) = x' = ct' = c(\alpha ct + \gamma t).$$

Therefore $\alpha = -\gamma v/c^2$. Our relationship now looks like

$$t' = \gamma(-vx/c^2 + t) \quad \text{and} \quad x' = \gamma(x - vt).$$

All that's left to do is deduce γ. To do this, let A shine another light pulse but this time 90 degrees away from the direction of the train, let's say this is in the y direction. According to Einstein, both A and B see the light pulse move away at a speed c. According to A, the position of the light pulse is given by $x = 0$ and $y = ct$. According to B, the light pulse travels in both the x' and y' direction, so by the Pythagorean theorem, $x'^2 + y'^2 = (ct')^2$. Therefore

$$\gamma^2(0 - vt)^2 + (ct)^2 = x'^2 + y'^2 = c^2t'^2 = c^2\gamma^2\big(-(v/c^2)0 + t\big)^2.$$

Solving for γ gives $\gamma = \pm 1/\sqrt{1 - v^2/c^2}$. We take the positive square root; otherwise, when $v = 0$, we would get $x' = -x$ rather than $x' = x$. To summarize, the correct transformation law between (t', x') and (t, x) is

$$t' = \frac{t - vx/c^2}{\sqrt{1 - v^2/c^2}} \quad \text{and} \quad x' = \frac{x - vt}{\sqrt{1 - v^2/c^2}}.$$

How does this all this relate to the existence of a Lorentz metric on our manifold? (t, x, y, z) and (t', x', y', z') are merely coordinates used on the manifold, so we seek a

quantity which is coordinate independent. Using the transformation law just derived, we find

$$\begin{aligned}
-(ct')^2 + x'^2 + y'^2 + z'^2 &= -\gamma^2 c^2 (t - vx/c^2)^2 + \gamma^2 (x - vt)^2 + y^2 + z^2 \\
&= \frac{1}{1 - v^2/c^2} \left[-c^2 t^2 \left(1 - \frac{v^2}{c^2} \right) + x^2 \left(1 - \frac{v^2}{c^2} \right) \right] + y^2 + z^2 \\
&= -c^2 t^2 + x^2 + y^2 + z^2.
\end{aligned}$$

This is an invariant quantity on the manifold which does *not* depend on the coordinates used to describe it. But this is precisely the quantity of a nondegenerate, quadratic form g with signature $(-,+,+,+)$ applied to the vector

$$\begin{aligned}
v &= t\frac{\partial}{\partial t} + x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y} + z\frac{\partial}{\partial z} \\
&= t'\frac{\partial}{\partial t'} + x'\frac{\partial}{\partial x'} + y\frac{\partial}{\partial y} + z\frac{\partial}{\partial z}
\end{aligned}$$

provided that these are orthogonal bases for the tangent space satisfying

$$g\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial x}\right) = g\left(\frac{\partial}{\partial x'}, \frac{\partial}{\partial x'}\right) = g\left(\frac{\partial}{\partial y}, \frac{\partial}{\partial y}\right) = g\left(\frac{\partial}{\partial z}, \frac{\partial}{\partial z}\right) = 1$$

and

$$g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) = g\left(\frac{\partial}{\partial t'}, \frac{\partial}{\partial t'}\right) = -c^2$$

In conclusion, Einstein believed that the speed of light is measured to be c in any observer's reference frame. From this we were led to a new way of relating space and time coordinates between different reference frames. This relation allowed us to find a quantity which was invariant on the manifold (i.e. it did not depend on coordinates), and we found that this quantity is exactly described by a Lorentzian metric.

2 Spacetime

A ***spacetime*** is a Hausdorff, connected, second countable smooth manifold M endowed with a smooth nondegenerate Lorentzian metric g with signature $(-,+,+,+)$. A vector $v \in T_pM$ is said to be ***timelike***, ***null***, or ***spacelike*** if $g(v,v)$ is negative, zero, or positive, respectively. Likewise, embedded submanifolds are said to be timelike, null, or spacelike if every tangent vector on the submanifold is timelike, null, or spacelike, respectively. The set of null vectors in T_pM defines the ***lightcone*** at $p \in M$. Timelike vectors are within the lightcone and spacelike vectors are outside the lightcone. A piecewise smooth curve γ is timelike, null, or spacelike if the tangent vector γ' is timelike, null, or spacelike where defined along γ.

Physically, timelike curves are curves whose velocities are "traveling slower than light". Null curves are curves that are "traveling at the speed of light", so they can physically represent particles like photons. A spacelike curve would be one that is "traveling faster than light." If γ is a timelike curve parametrized by s, then the ***proper time*** τ of γ is defined by $\tau = \frac{1}{c}\int \sqrt{-g(\gamma',\gamma')}ds$ (the integral is taken over the intervals where γ' is defined) or we can use the Lebesgue integral. If an observer is following the trajectory of γ, then τ measures the amount of time that particular observer experiences.

Example: Minkowski Space

Minkowski space is the spacetime with manifold $\mathbb{R}^4$ and a flat Lorentz metric η. This is the spacetime with no gravitational effects. An observer moving with speed v in Minkowski space will continue to move with speed v indefinitely. This is precisely what

we expect from an observer whose path is far away from any massive bodies like the Earth or Sun. Using coordinates (t, x, y, z), the metric can be written as

$$\eta = -c^2dt^2 + dx^2 + dy^2 + dz^2.$$

where c is the speed of light. Physically, these coordinates correspond to an inertial reference frame. One can imagine setting meter sticks along the x, y, and z axes and a clock at each point of (x, y, z). In these coordinates, imagine an observer A moves through $\mathbb{R}^4$ along the curve $\gamma_A(s) = (s, 0, 0, 0)$ for $s \in (0, 1)$ (i.e. this observer is not moving). The proper time for observer A is

$$\tau_A = \frac{1}{c}\int_0^1 \sqrt{-\eta(\gamma_A', \gamma_A')}ds = \frac{1}{c}\int_0^1 \sqrt{c^2}ds = 1.$$

Now let's consider an observer B who moves at a speed $v < c$ in the x direction relative to observer A. The path of observer B is $\gamma_B(s) = (s, vs, 0, 0)$. Its proper time is

$$\tau_B = \frac{1}{c}\int_0^1 \sqrt{-\eta(\gamma_B', \gamma_B')}ds = \frac{1}{c}\int_0^1 \sqrt{c^2 - v^2}ds = \sqrt{1 - v^2/c^2}.$$

We see that $\tau_B < \tau_A$. This means that observer A experienced more time than observer B. This phenomenon is known as *time dilation* and has the slogan "moving clocks run slow." However according to observer B, A is the one that is moving so B will see A's clock running slower. The solution to this paradox is that "time running slower" is a relative concept. It depends on the observer (i.e. reference frame, coordinate system, etc.) that your working with. "time" is *not* an invariant quantity.

Let us consider all possible timelike paths from $(0, 0, 0, 0)$ to $(1, 0, 0, 0)$. It is easy to convince yourself that the path which maximizes the proper time of an observer is precisely the "straightest" path $\gamma(s) = (s, 0, 0, 0)$. This is a peculiar quirk. Observers

who move with constant velocity are moving along paths which maximizes their proper time.

So far we haven't considered gravitational effects. To find a description of gravity let's imagine an observer inside a box with no windows traveling in space. This observer will not be able to deduce if he is floating in free space (i.e. Minkowski space) or is in orbit around the sun. This is known as the ***equivalence principle***. A classic example is astronauts aboard the international space station. These astronauts are pulled by Earth's gravity, but if they had no windows, then the astronauts would be unable to know if they were orbiting the Earth, falling towards the Sun, or just floating in free space (i.e. as straight timelike curves in Minkowski space). This is because in Newtonian Mechanics the observer would feel a force

$$\vec{F} = m\vec{a} = \frac{GM_{\text{Sun}}m}{r^2}\hat{r}.$$

The mass of the observer m cancels and so $\vec{a}$ has no dependence on m. This means the fictitious force the observer feels in the box exactly cancels the force felt by gravity. However we know that Newton's description of gravity is incorrect because it allows objects to be accelerated faster than the speed of light. Moreover, there is no coordinate independent description of Newtonian gravity.

Perhaps the observer in the box doesn't know the difference between orbiting the sun and floating in free space because both paths have the same defining property. But what property? We already know that observers who move with constant velocities in Minkowski space are moving along the straightest paths which maximizes their proper time. Perhaps the observer in the box orbiting the sun is also moving along a path

which (locally) maximizes his proper time and the affects of gravity are merely what he perceives from following this special path.

3 Covariant Differentiation and Geodesics

In Riemannian Geometry geodesics are the curves which locally minimize their length. In the same way we will see that timelike geodesics locally maximize their proper time. In this section we develop the machinery to show this. The tools developed here (affine connection and parallel transport) are no different than the ones used to study Riemannian geometry.

3.1 Covariant Differentiation

We seek a way to differentiate vector fields on our manifold which is independent of coordinates. Let M be a smooth manifold. A ***derivative operator*** (or ***affine connection***) ∇ is a rule which assigns to each field field v a differential operator ∇_v which maps an arbitrary vector field w into another vector field $\nabla_v w$ that satisfies the following three properties:

(1) $\nabla_{fv+u} w = f\nabla_v w + \nabla_u w$;

(2) $\nabla_v(u + w) = \nabla_v u + \nabla_v w$;

(3) $\nabla_v(fw) = f\nabla_v w + v(f)w$.

for any smooth function f and smooth vector fields u, v, and w.

We say that $\nabla_v w$ is the ***covariant derivative*** of w the direction v with respect to ∇. We will also write ∇w for the map $v \mapsto \nabla_v w$. Therefore property (3) is equivalent to $\nabla(fw) = df \otimes w + f\nabla w$. Suppose $\{e_a\}$ is a vector basis with dual one-form basis $\{e^a\}$ on a neighborhood U of M. If the components of v and w with respect to $\{e_a\}$ are $\{v^a\}$ and $\{w^a\}$, then we write the components of $\nabla_v w$ as $v^b \nabla_b w^a$ and the components of ∇w as $\nabla_a v^b$, so

$$\nabla w = (\nabla_b w^a) e^b \otimes e_a.$$

By the three properties, ∇ is completely determined by the smooth functions $\Gamma^a{}_{bc}$ defined by

$$\nabla e_c = \Gamma^a{}_{bc} e^b \otimes e_a \quad \text{which is equivalent to} \quad \Gamma^a{}_{bc} = e^a\big(\nabla_{e_b} e_c\big).$$

Therefore

$$\nabla w = \nabla(w^c e_c) = dw^c \otimes e_c + w^c \Gamma^a{}_{bc} e^b \otimes e_a.$$

If we consider a coordinate basis $\{e_a\} = \{\partial/\partial x^a\}$, then the components of ∇w are

$$\nabla_b w^a = \frac{\partial w^a}{\partial x^b} + \Gamma^a{}_{bc} Y^c.$$

For a coordinate basis, the smooth functions $\Gamma^a{}_{bc}$ are known as the ***Christoffel symbols***.

We can extend the definition of a covariant derivative to any smooth tensor field by the following rules:

(4) if T is a smooth tensor field of type (p, q) (i.e. it takes in p covectors and q vectors), then ∇T is a smooth tensor field of type $(p, q+1)$;

(5) ∇ is linear: if T and S are smooth tensor fields of type (p, q), then $\nabla(\alpha T + S) = \alpha \nabla T + \nabla S$ for any real number α;

(6) ∇ commutes with contractions;

(7) ∇ obeys a Leibniz rule: if T is a smooth tensor field of type (p, q) and S is a smooth tensor field of type (p', q'), then $\nabla(S \otimes T) = \nabla S \otimes T + S \otimes \nabla T$;

(8) $\nabla f = df$ for any smooth real-valued function f.

Given a basis, we write the components of ∇T as $\nabla_c T^{a_1 \cdots a_p}{}_{b_1 \cdots b_q}$ where $T^{a_1 \cdots a_p}{}_{b_1 \cdots b_q}$ are the components of T with respect to the basis. By properties (6) and (7), we have

$$\begin{aligned} 0 &= \nabla_{e_b}\big(e^a(e_c)\big) \\ &= \nabla_{e_b} e_c \otimes e^a + \nabla_{e_b} e^a \otimes e_c \\ &= e^a(\nabla_{e_b} e_c) + e_c(\nabla_{e_b} e^a) \\ &= \Gamma^d{}_{bc}\delta^a{}_d + e_c(\nabla_{e_b} e^a) \end{aligned}$$

Therefore $\nabla_{e_b} e^a = -\Gamma^a{}_{bc} e^c$. So if we consider a coordinate basis $\{\partial/\partial x^a\}$ and its dual basis $\{dx^a\}$, the components of ∇T can be computed using the Christoffel symbols:

$$\begin{aligned} \nabla_c T^{a_1 \cdots a_p}{}_{b_1 \cdots b_q} =& \frac{\partial T^{a_1 \cdots a_p}{}_{b_1 \cdots b_q}}{\partial x^a} + \Gamma^{a_1}{}_{cd} T^{d a_2 \cdots a_p}{}_{b_1 \cdots b_q} + \Gamma^{a_2}{}_{cd} T^{a_1 d a_3 \cdots a_p}{}_{b_1 \cdots b_q} + \cdots + \Gamma^{a_p}{}_{cd} T^{a_1 \cdots a_{p-1} d}{}_{b_1 \cdots b_q} \\ & - \Gamma^d{}_{cb_1} T^{a_1 \cdots a_p}{}_{d b_2 \cdots b_q} - \Gamma^d{}_{cb_2} T^{a_1 \cdots a_p}{}_{b_1 d b_3 \cdots b_q} - \cdots - \Gamma^d{}_{cb_q} T^{a_1 \cdots a_p}{}_{b_1 \cdots b_{q-1} d}. \end{aligned}$$

Thus a knowledge of the Christoffel symbols determines the affine connection ∇.

3.2 Parallel Transport and Geodesics

In our Minkowski space example, we noticed that the curve which maximized proper time between two points was the straight line between those two points. We know what it means for a curve to be straight in $\mathbb{R}^4$ but we have to adopt a definition for an arbitrary

smooth manifold M. If T is a smooth tensor field defined along a smooth curve $\gamma(s)$, we define $DT/\partial s$, as the ***covariant derivative of*** T ***along*** γ, as $\nabla_{\partial/\partial s}\tilde{T}$ where $\tilde{T}$ is any tensor field T extending T onto an open neighborhood of γ. One can show $DT/\partial s$ is independent of the extension (see [2] and [5]).

T is said to be ***parallelly transported along*** γ if $DT/\partial s = 0$. Given a smooth curve γ with endpoints p and q and a tensor defined at p, the theory of solutions of ordinary differential equations guarantees a unique tensor at q by parallelly transferring the tensor from p along γ. If $\gamma : [0, 1] \to \mathbb{R}^n$ is a smooth curve in $\mathbb{R}^n$ with the derivative operator given by regular differentiation and v is a vector at $\gamma(0)$, then the curve which is traced by the parallel transported vector v is parallel (in the usual sense) to the curve γ.

The curve $\gamma(s)$ is said to be a ***geodesic*** if one can find a parametrization $\phi(s)$ such that

$$\frac{D}{d\phi}\left(\frac{\partial}{\partial\phi}\right)\Bigg|_{\gamma} = 0.$$

In this case ϕ is called an ***affine parameter***. If s is already an affine parameter, then $\nabla_{\gamma'}\gamma' = 0$. Thus a geodesic is a curve whose tangent vector is parallelly transported along itself. The geodesics in $\mathbb{R}^n$ are precisely the straight lines in $\mathbb{R}^n$.

Suppose $\gamma(s)$ is a geodesic and s is an affine parameter and $\{x^a\}$ are coordinates about some points of γ and and $\gamma(s)$ has coordinates $\{x^a(s)\}$, then $\nabla_{\gamma'}\gamma' = 0$ is equivalent to the equation

$$\frac{d^2x^a}{ds^2} + \Gamma^a{}_{bc}\frac{dx^b}{ds}\frac{dx^c}{ds} = 0.$$

The above equation is known as the ***geodesic equation***. Notice that if s is an affine parameter, then ϕ is an affine parameter if and only if $\phi(s) = as + b$ for some numbers a and b. The existence and uniqueness theorems for ordinary differential equations applied to the geodesic equation show that for any point $p \in M$ and any vector v at p, there exists a unique maximal geodesic $\gamma_v(s)$ starting at p and initial direction v. Therefore we can define a smooth map $\exp_p$, called the ***exponential map at*** p, from a subset of T_pM to M, where for each $v \in T_pM$, $\exp_p(v)$ is the point in M a unit parameter distance along the geodesic γ_v from p. $\exp_p$ may not be defined for all $v \in T_pM$, since the geodesic $\gamma_v(s)$ may not be defined for all s (e.g. if M is $\mathbb{R}^4$ with a point removed). If s does take all values in $\mathbb{R}$, the geodesic $\gamma(s)$ will be said to be a ***complete*** geodesic. The manifold M is said to be ***geodesically complete*** if all geodesics on M are complete, that is if $\exp_p$ is defined on all T_pM for every point $p \in M$. The singularity theorems prove existence of incomplete geodesics. For timelike geodesics this means time has a beginning or time has an end for the observer following such an unfortunate geodesic.

The differential $(d\exp_p)_0$ is the identity on T_pM, so it follows from the inverse function theorem that $\exp_p$ is a local diffeomorphism. If $\exp_p : N_0 \to N_p$ is a local diffeomorphism, then N_p is said to be ***normal neighborhood*** of p. In fact, N_p can be chosen to be ***convex***, i.e. for any $q, r \in N_p$ there is a unique geodesic, γ, completely contained in N_p, which joins q and r (see [2] and [5] and note that their proofs don't rely on the signature of the metric). In a convex normal neighborhood N_p, one can define ***normal coordinates*** $\{x^a\}$ by choosing any point $q \in N_p$, choosing a basis $\{e_a\}$ of T_q, and defining the coordinates of the point $r \in N_p$ via $r = \exp(x^a e_a)$, e.g. the coordinates are taken from the natural coordinates on T_qM. Then $\partial/\partial x^a\big|_q = e_a$, and by the geodesic

equation, we have $(\Gamma^a{}_{bc} + \Gamma^a{}_{cb})\big|_q = 0$.

Let v and w be smooth vector fields. The ***Lie derivative*** of w with respect to v is the smooth vector field $[v, w]$ defined by $[v, w](f) = v\big(w(f)\big) - w\big(v(f)\big)$ for all smooth functions f. Given a derivative operator ∇, the ***torsion tensor*** is a $(1, 2)$ tensor field T defined by $T(v, w) = \nabla_v w - \nabla_w v - [v, w]$. Using a coordinate basis $\{\partial/\partial x^a\}$, its components are given by $T^a{}_{bc} = \Gamma^a{}_{bc} - \Gamma^a{}_{cb}$. We will only be working with ***torsion-free*** derivative operators, i.e. $T = 0$. This means $\Gamma^a{}_{bc} = \Gamma^a{}_{cb}$, so when using normal coordinates at $p \in M$, we have $\Gamma^a{}_{bc}\big|_p = 0$. One useful property of torsion-free connections is that $\nabla_a \nabla_b f = \nabla_b \nabla_a f$ with respect to any basis. To see this, simply expand $\nabla_v w(f) - \nabla_w(f) = [v, w](f)$ in terms of coordinates. We have

$$\begin{aligned}
\nabla_v w(f) - \nabla_w v(f) &= [v, w](f) \\
&= v^a \nabla_a (w^b \nabla_b f) - w^b \nabla_b (v^a \nabla_a f) \\
&= v^a (w^b \nabla_a \nabla_b f + \nabla_b f \nabla_a w^b) - w^b (v^a \nabla_b \nabla_a f + \nabla_a f \nabla_b v^a) \\
&= \nabla_v w(f) - \nabla_w v(f) + v^a w^b (\nabla_a \nabla_b f - \nabla_b \nabla_a f).
\end{aligned}$$

Hence $\nabla_a \nabla_b f - \nabla_b \nabla_a f = 0$ for all smooth functions. Conversely, assuming $\nabla_a \nabla_b f - \nabla_b \nabla_a f = 0$ for all smooth functions f implies ∇ is torsion-free.

3.3 The Metric

Now let us suppose M has a ***metric*** g on it, i.e. a smooth, symmetric tensor of type (0,2) that is non-degenerate. Vectors v and w are ***orthogonal*** if $g(v, w) = 0$. In a vector basis $\{e_a\}$ with dual one-form basis $\{e^a\}$, the components of g will be written as

g_{ab}. The ***signature*** of g is the pair (p, q) where p is the number of negative eigenvalues of the matrix (g_{ab}) and q is the number of positive eigenvalues of the matrix (g_{ab}). g is ***Lorentzian*** if $p = 1$ and $q \geq 1$. g is ***Riemannian*** if $p = 0$ and $q \geq 1$. Since g is nondegenerate, it has an inverse g^{-1} which is a smooth, symmetric tensor of type (2,0) with components g^{ab} that satisfy $g^{ab}g_{bc} = \delta^a{}_c$, i.e. it's the identity map from the tangent space to itself. Given a vector field v with components v^a, the metric induces a natural covector field with components $v_a = v^b g_{ab}$. Likewise, given any covector field with components ω_a, the metric induces a natural vector field with components $\omega^a = \omega_b g^{ab}$. This process is called ***lowering and raising the index***, respectively, and it can be applied to any tensor field of any type, i.e if S is a (2,1) tensor with components $S^{ab}{}_c$, then we can define a $(1, 2)$ tensor with components $S^a{}_{bc} = S^{ad}{}_c g_{db}$.

Given a smooth curve γ in $\mathbb{R}^n$ and vectors v and w which are parallelly propagated along γ, their inner product $v \cdot w = g_{\text{Euclid}}(v, w)$ is constant along the curve. We can capture this notion in the setting of smooth manifolds with metrics. Suppose M is a smooth manifold with a smooth metric g and γ is a curve in M beginning at p and ending at q. Suppose v and w are vectors fields that are parallelly propagated along γ. Then if we want the $g(v, w)$ to be constant along the curve, we want $\gamma'\big(g(v, w)\big) = 0$. By property (8) of the covariant derivative, this is equivalent to $\nabla_{\gamma'} g(v, w) = 0$. If $\{e_a\}$ is any vector basis, then by properties (6) and (7) of the covariant derivative, this is equivalent to

$$0 = \nabla_{\gamma'} g(v, w) = v^a w^b \nabla_{\gamma'} g_{ab} + g_{ab} \nabla_{\gamma'} (v^a w^b .)$$

The last term is zero since v and w are parallelly propagated, therefore we desire $v^a w^b \nabla_{\gamma'} g_{ab} = 0$. Since we want this to hold for any parallelly propagated vectors v and w and any loops γ, we want the derivative operator to satisfy $\nabla g = 0$, i.e. $\nabla_c g_{ab} = 0$. If this is the case, we say that ∇ is ***compatible*** with g.

Theorem 3.1 *If M is a smooth manifold with smooth metric g, then there exists a unique derivative operator ∇ such that ∇ is torsion free and compatible with g.*

Proof of Theorem 3.1. Recall that a derivative operator is completely determined by its Christoffel symbols so it suffices to work in a coordinate neighborhood $\{x^a\}$. First suppose ∇ exists, then the components of ∇g with respect to $\{\partial/\partial x^a\}$ are given by

$$0 = \nabla_a g_{bc} = \frac{\partial g_{bc}}{\partial x^a} - \Gamma^d{}_{ab} g_{dc} - \Gamma^d{}_{ac} g_{bd}.$$

Therefore $\Gamma_{cab} + \Gamma_{bac} = \partial g_{bc}/\partial x^a$. Cyclic permuting the indices, we also have $\Gamma_{cba} + \Gamma_{abc} = \partial g_{ac}/\partial x^b$ and $\Gamma_{bca} + \Gamma_{acb} = \partial g_{ab}/\partial x^c$. Since ∇ is torsion free (i.e. $\Gamma_{abc} = \Gamma_{acb}$), adding the first two equations and subtracting the third yields

$$\Gamma_{cab} = \frac{1}{2}\left(\frac{\partial g_{bc}}{\partial x^a} + \frac{\partial g_{ac}}{\partial x^b} - \frac{g_{ab}}{\partial x^c}\right).$$

Thus, if we choose our Christoffel symbols to satisfy the above equation, then ∇ is uniquely determined and is automatically torsion free and compatible with g. □

From now on, we will only be working with the unique derivative operator ∇ determined by Theorem 3.1. Notice that since the Christoffel symbols are defined in terms of derivatives of the metric components, it follows that the unique derivative operator

for Minkowski space, $(\mathbb{R}^4, \eta)$, is that of ordinary partial differentiation. An immediate consequence of the compatibility of the metric is the following proposition:

Proposition 3.2 *Suppose $\gamma(s)$ is a geodesic, then $g\big(\gamma'(s), \gamma'(s)\big)$ is constant along γ.*

Proof. By the symmetry and compatibility of the metric, we have

$$\frac{d}{ds} g(\gamma', \gamma') = \gamma' g(\gamma', \gamma') = 2g(\nabla_{\gamma'}\gamma', \gamma') = 0$$

since $\nabla_{\gamma'}\gamma' = 0$. □

Corollary 3.3 *Timelike, null, and spacelike geodesics remain timelike, null, and spacelike.*

3.4 Timelike Geodesics Maximize Proper Time

Let (M, g) be a spacetime. Motivated by the equivalence principle in section 2, we seek curves that locally maximize their proper time. In this section, we will show that timelike geodesics are precisely these curves. The following lemma is an analogue of the Gauss lemma in Reimannian Geometry.

Lemma 3.4 *Let N_p be a normal neighborhood of a point $p \in M$ and $f : N_p \to \mathbb{R}$ defined by $f(q) = g(\exp_p^{-1} q, \exp_p^{-1} q)$. Then the timelike geodesics through p are orthogonal to the three-surfaces of constant, negative f. In other words, the surfaces of constant f are spacelike.*

Proof. Let $v(r)$ denote the tangent to a curve in N_p, where $g\big(v(r), v(r)\big) = -c^2$. Define the curves $\lambda(r) = \exp_p\big(s_0 v(r)\big)$ with s_0 constant and small enough so λ is defined. We want to show that the timelike geodesics $\gamma(s) = \exp_p\big(sv(r_0)\big)$ (with r_0 constant) are orthogonal to the curves $\lambda(r)$. So in terms of the two-surface $\alpha(s,r) = \exp_p\big(sv(r)\big)$, we want to show $h(s,r) = g\left(\partial/\partial s\big|_{\alpha(s,r)}, \partial/\partial r\big|_{\alpha(s,r)}\right) = 0$ where we are denoting $\partial/\partial s\big|_\alpha$ as the push forward of $\partial/\partial s$ under α and likewise with $\partial/\partial r\big|_\alpha$. Since ∇ is compatible with g, we find

$$\frac{\partial}{\partial s}h = g\left(\frac{D}{\partial s}\frac{\partial}{\partial s}\bigg|_\alpha, \frac{\partial}{\partial r}\bigg|_\alpha\right) + g\left(\frac{\partial}{\partial s}\bigg|_\alpha, \frac{D}{\partial s}\frac{\partial}{\partial r}\bigg|_\alpha\right).$$

The first term is zero since γ is a geodeisc. Now since ∇ is torsion-free and s, r are coordinates of a two-dimensional surface, we have $\frac{D}{\partial s}\frac{\partial}{\partial r} = \frac{D}{\partial r}\frac{\partial}{\partial s}$ (i.e. their Lie derivative is zero). Therefore

$$\frac{\partial}{\partial s}h = g\left(\frac{\partial}{\partial s}\bigg|_\alpha, \frac{D}{\partial r}\frac{\partial}{\partial s}\bigg|_\alpha\right) = \frac{1}{2}\frac{\partial}{\partial r}g\left(\frac{\partial}{\partial s}\bigg|_\alpha, \frac{\partial}{\partial s}\bigg|_\alpha\right) = \frac{\partial}{\partial r}(-c^2) = 0.$$

Therefore h is independent of s, but $h(0,r) = 0$ since $\partial/\partial r\big|_{\alpha(0,r)} = 0$. Thus h is identically zero. □

The next proposition is physically intuitive but deceptively difficult to prove. We will use it countless times when we discuss causality in chapter 6. The timelike curves in the following proposition and theorem can assumed to be continuous and piecewise smooth, but at any point the curve is not differentiable, the left and right tangent vectors both point within the same half of the lightcone.

Proposition 3.5 *Let N_p be a convex normal neighborhood of a point $p \in M$. Then the points $q \in N_p$ which can be reached from timelike (respectively, causal) curves in N_p are*

those of the form $\exp_p(v), v \in T_pM$ *where* $g(v,v) < 0$ *(respectively* ≤ 0*).*

Proof. We consider timelike curves first. Let C_p denote the set of all timelike vectors at p and suppose $\gamma(s)$ is a timelike curve in N_p. Initially γ is timelike, so it must enter $\exp_p(C_p)$. We need to show that γ remains in $\exp_p(C_p)$. Notice that $\exp_p(C_p) = \{q : f(q) < 0\}$ where f is defined in the previous Lemma. Since the surfaces of constant f are spacelike, f must decrease along γ since it'sγ is timelike and at any non-differentiable point the tangent vectors of γ point in the same half of the lightcone. Therefore γ must remain in $\{q : f(q) < 0\}$.

Now we prove the theorem for causal curves. Let $\gamma(s)$ be a causal curve in N_p. Initially, γ enters $\exp_p(\overline{C_p})$. We want to show γ remains in $\exp_p(\overline{C_p})$. The trick is to vary γ slightly making it into a timelike curve. Let v be a smooth vector field on T_pM and denote $\widetilde{v}$ as the push forward of v from the exponential map. Construct v such that $\widetilde{v}$ is everywhere timelike and $g\big(\widetilde{v}(p), \gamma'(p)\big) < 0$ (i.e. $\widetilde{v}(p)$ and $\gamma'(p)$ point in the same half of the light cone). Let $\overline{\gamma}(s) = \exp_p^{-1}\big(\gamma(s)\big)$. Now for each $\epsilon \geq 0$, we define the curve $\beta_\epsilon(s)$ in T_pM by demanding $\beta'_\epsilon(s) = \overline{\gamma}'(s) + \epsilon v\big|_{\beta_\epsilon(s)}$. We see that for each $\epsilon > 0$, $\exp_p\big(\beta_\epsilon(s)\big)$ is a timelike curve in N_p and so is contained in $\exp_p(C_p)$ by the above paragraph. Thus the causal curve $\gamma(s) = \exp_p\big(\beta(s,0)\big)$ is contained in $\overline{\exp_p(C_p)} = \exp_p(\overline{C_p})$. $\square$

Now we can state the theorem which says that timelike geodesics are the unique curves which locally maximize proper time. Recall that the proper time τ of a timelike curve γ is defined to be $\tau = \frac{1}{c}\int \sqrt{-g(\gamma'(s),\gamma'(s))}ds$, and it physically represents that the amount of time an observer following the timelike curve γ experiences.

Theorem 3.6 *Let N_p be a convex normal neighborhood about a point $p \in M$. Let $q \in N_p$. If γ is the unique timelike geodesic connecting p to q, then $\tau_\gamma > \tau_\lambda$ where λ is any other smooth piecewise timelike curve connecting p to q.*

Proof. As in the lemma, let $\alpha(s,r) = \exp_p\big(sv(r)\big)$ where $g\big(v(r),v(r)\big) = -c^2$. We can uniquely write the curve λ as $\lambda(r) = \alpha\big(h(r),r\big)$ where h is some continuous and piecewise smooth function. By the chain rule, we get

$$\lambda' = h' \frac{\partial}{\partial s}\bigg|_\alpha + \frac{\partial}{\partial r}\bigg|_\alpha .$$

By the lemma, $g\left(\partial/\partial s\big|_\alpha, \partial/\partial r\big|_\alpha\right) = 0$ and $\partial/\partial r|_\alpha$ is either spacelike or the zero vector. So since $g\left(\partial/\partial s\big|_\alpha, \partial/\partial s\big|_\alpha\right) = -1$ and $g\left(\partial/\partial r\big|_\alpha, \partial/\partial r\big|_\alpha\right) \geq 0$, we have

$$g\left(\frac{\partial}{\partial r}\bigg|_\lambda, \frac{\partial}{\partial r}\bigg|_\lambda\right) = -|h'(r)|^2 + g\left(\frac{\partial}{\partial r}\bigg|_\alpha, \frac{\partial}{\partial r}\bigg|_\alpha\right) \geq -|h'(r)|^2.$$

Equality holds if and only if $\partial/\partial r\big|_\alpha = 0$, i.e. if and only if λ is a timelike geodesic. Therefore

$$\tau_\lambda \leq \frac{1}{c}\int h'(r)dr = \tau_\gamma,$$

with equality if and only if λ is a timelike geodesic. □

Thus we have shown that timelike geodesics in a spacetime are the paths which observers locally maximize their proper time. We can actually prove theorem 3.6 quickly using what is known as a synchronous coordinate system. Since this type of coordinate system will be useful when we talk about congruences in section 7.1, we will introduce them here while proving theorem 3.6.

Alternate Proof of Theorem 3.6. Extend γ so that we can consider a point $r \in \gamma$ such that r comes before p and q on γ. Choose normal coordinates (t, x, y, z) for N with origin at r such that the light cone in T_rM is defined by $c^2t^2 = x^2 + y^2 + z^2$. In the region $ct > \sqrt{x^2 + y^2 + z^2}$ let us construct new coordinates (T, X, Y, Z) by

$$cT = \sqrt{(ct)^2 - x^2 - y^2 - z^2}$$

$$X = \frac{x}{t}, \;\; Y = \frac{y}{t}, \;\; Z = \frac{z}{t}.$$

Then timelike geodesics emanating from r are described by the curves $X, Y, Z = \text{const}$ and are orthogonal to the spacelike hypersurfaces $T = \text{const}$. Thus what we have constructed is a ***synchronous coordinate system*** (i.e. normal coordinates in which constant spatial coordinates are timelike geodesics orthogonal to a system of spacelike coordinate hypersurfaces). Therefore there exists a positive definite symmetric matrix h_{ij}, which depends only on the coordinates $\{X, Y, Z\}$, such that the metric takes the form

$$g = -c^2 dT^2 + h_{ij} dX^i dX^j.$$

Let s be the parameter for any piecewise smooth timelike curve λ. Then the proper time of λ is

$$\tau_\lambda = \frac{1}{c} \int_a^b \sqrt{c^2 - h_{ij} \frac{dX^i}{ds} \frac{dX^j}{ds}} ds.$$

Any timelike curve connecting p to q which is not a geodesic will have nonzero components $\frac{dX^i}{ds}$ on a set with positive measure whereas a timelike geodesic will have $\frac{dX^i}{ds} = 0$ everywhere. $\square$

Motivated by the equivalence principle, we established that observers in a spacetime move on timelike geodesics. But how does this notion reconcile with the familiar

gravitational laws of Newton? We will see that it's the curvature tensor which produces gravitational effects.

4 Gravity as Curvature

4.1 Riemann Curvature Tensor

Let M be a smooth manifold with any metric g. Given smooth vectors fields u, v, w, the ***Riemann curvature tensor*** is a smooth vector field $R(u,v)w$ defined by

$$R(u,v)w = \nabla_u(\nabla_v w) - \nabla_v(\nabla_u w) - \nabla_{[u,v]}w.$$

The fact that the Riemann curvature tensor is indeed a (3,1) tensor can be checked by direct computation. Let $\{e_a\}$ be a vector basis with dual one-form basis $\{e^a\}$, then by properties (6), (7), and (8) of ∇, we find

$$\begin{aligned}\nabla_u(\nabla_v w) &= \nabla_u\big(v^c\nabla_c(w^a e_a)\big)\\ &= v^c\nabla_u\big(\nabla_c(w^a e_a)\big) + u(v^c)\nabla_c(w^a e_a)\\ &= v^c u^b \nabla_b\nabla_c w^a e_a + u(v^c)\nabla_c(w^a e_a)\end{aligned}$$

Likewise, $\nabla_v(\nabla_u w) = u^c v^b \nabla_b\nabla_c w^a e_a + v(u^c)\nabla_c(w^a e_a)$. Thus, if the components of the Riemann tensor are given by $R^a{}_{bcd} = e^a\big(R(e_c, e_d)e_b\big)$, then the components of $R(u,v)w$

are given by

$$\begin{aligned} R^a{}_{bcd}u^c v^d w^b &= v^c u^b \nabla_b \nabla_c w^a + u(v^c)\nabla_c w^a - u^c v^b \nabla_b \nabla_c w^a - v(u^c)\nabla_c w^a - [u,v]^b \nabla_b w^a \\ &= v^c u^b \nabla_b \nabla_c w^a - u^c v^b \nabla_b \nabla_c w^a \\ &= v^c u^b (\nabla_b \nabla_c w^a - \nabla_c \nabla_b w^a) \\ &= v^d u^c (\nabla_c \nabla_d w^a - \nabla_d \nabla_c w^a). \end{aligned}$$

Since u and v were arbitrary vector fields, we see that

$$R^a{}_{bcd}w^b = \nabla_c \nabla_d w^a - \nabla_d \nabla_c w^a.$$

Here we specifically see the non-commutativity of the second covariant derivatives of w expressed in terms of the Riemann tensor. If the vector basis comes from a coordinate system $\{x^a\}$, then we can compute the components of the curvature tensor in terms of the Christoffel symbols and its derivatives. We have

$$\begin{aligned} \nabla_c \nabla_d w^a &= \frac{\partial}{\partial x^c}(\nabla_d w^a) - \Gamma^e{}_{cd}\nabla_e w^a - \Gamma^a{}_{cb}\nabla_d w^b \\ &= \frac{\partial^2 w^a}{\partial x^c \partial x^d} + \frac{\partial \Gamma^a{}_{db}}{\partial x^c}w^b + \Gamma^a{}_{db}\frac{\partial w^b}{\partial x^c} - \Gamma^e{}_{db}\frac{\partial}{\partial x^e}w^a - \Gamma^e{}_{cd}\Gamma^a{}_{eb}w^b + \Gamma^a{}_{cb}\frac{\partial w^a}{\partial x^d} + \Gamma^a{}_{ce}\Gamma^e{}_{db}w^b. \end{aligned}$$

Now using the torsion-free property and the fact that mixed partial derivatives commute, we find

$$\begin{aligned} R^a{}_{bcd}w^b &= \nabla_c \nabla_d w^a - \nabla_d \nabla_c w^a \\ &= \left(\frac{\partial \Gamma^a{}_{db}}{\partial x^c} - \frac{\partial \Gamma^a{}_{cb}}{\partial x^d} + \Gamma^a{}_{ce}\Gamma^e{}_{db} - \Gamma^a{}_{de}\Gamma^e{}_{cb} \right) w^b \end{aligned}$$

Since this expression is true for all vectors w, we have the following coordinate expression for the components of the curvature tensor

$$R^a{}_{bcd} = \frac{\partial \Gamma^a{}_{db}}{\partial x^c} - \frac{\partial \Gamma^a{}_{cb}}{\partial x^d} + \Gamma^a{}_{ce}\Gamma^e{}_{db} - \Gamma^a{}_{de}\Gamma^e{}_{cb}.$$

Theorem 4.1 *The curvature tensor has the following symmetries which we express using its components in any vector basis* $\{e_a\}$:

(1) $R^a{}_{bcd} = -R^a{}_{bdc}$

(2) $R^a{}_{bcd} + R^a{}_{dbc} + R^a{}_{cdb} = 0.$

(3) $R_{abcd} = -R_{bacd}$

(4) $R_{abcd} = R_{cdab}.$

(5) $\nabla_e R^a{}_{bcd} + \nabla_d R^a{}_{bec} + \nabla_c R^a{}_{bde} = 0.$

Proof. (1) follows from definition. (2) follows from the Jacobi identity, i.e. $\big[u, [v, w]\big] + \big[w, [u, v]\big] + \big[v, [w, u]\big] = 0$, and the fact that ∇ is torsion-free. (3) is equivalent to the statement $g(R(u, v)w, w) = 0$ which follows from ∇ being compatible with g. (4) follows applying (2) and (3) to the sum of the four equations

$$R_{abcd} + R_{adbc} + R_{acdb} = 0$$

$$R_{bacd} + R_{adac} + R_{bcda} = 0$$

$$R_{cabd} + R_{cdab} + R_{cbda} = 0$$

$$R_{dabc} + R_{dcab} + R_{dbca} = 0.$$

It suffices to prove (5) at a point $p \in M$ since ∇R is a tensor. Let $\{e_a\}$ be a vector basis corresponding to a coordinate basis of normal coordinates at $p \in M$ with dual one-form basis $\{e^a\}$. So by linearity of ∇R, it suffices to show

$$\nabla R(e^a, e_b, e_c, e_d, e_e) + \nabla R(e^a, e_b, e_e, e_c, e_d) + \nabla R(e^a, e_b, e_d, e_e, e_c) = 0$$

at $p \in M$. By definition

$$\nabla R(e^a, e_b, e_c, e_d, e_e) = e^a\big(\nabla_{e_e} R(e_c, e_d)e_b\big) = e^a\big(\nabla_{e_e}\nabla_{e_c}\nabla_{e_d}e_b + \nabla_{e_e}\nabla_{e_d}\nabla_{e_c}e_b + \nabla_{e_e}\nabla_{[e_c,e_d]}e_b\big).$$

But since $\{e_a\}$ is the vector basis of a coordinate basis, $[e_a, e_b] = 0$. Therefore

$$\begin{aligned}
&\nabla R(e^a, e_b, e_c, e_d, e_e) + \nabla R(e^a, e_b, e_e, e_c, e_d) + \nabla R(e^a, e_b, e_d, e_e, e_c) \\
=&e^a\big(\nabla_{e_e}\nabla_{e_c}\nabla_{e_d}e_b + \nabla_{e_e}\nabla_{e_d}\nabla_{e_c}e_b + \nabla_{e_d}\nabla_{e_e}\nabla_{e_c}e_b \\
&\nabla_{e_d}\nabla_{e_c}\nabla_{e_e}e_b + \nabla_{e_c}\nabla_{e_d}\nabla_{e_e}e_b + \nabla_{e_c}\nabla_{e_e}\nabla_{e_d}e_b\big) \\
=&e^a\big(R(e_e, e_c)\nabla_{e_d}e_b + R(e_e, e_d)\nabla_{e_c}e_b + R(e_d, e_e)\nabla_{e_c}e_b\big) \\
=&0.
\end{aligned}$$

The last equality follows because since we're working in normal coordinates, $\Gamma^a_{\ bc}\big|_p = 0$, so $\nabla_{e_a}e_b = 0$ at p. □

From contracting the first and third indices of the Riemann tensor, we arrive at the ***Ricci tensor***, Ric, whose components are given by $R_{bd} = R^a_{\ bad}$. By property (4) of Theorem 3.2, Ric is symmetric: $R_{ab} = R_{ba}$. We also define the ***scalar curvature***, R (unfortunately the same symbol used for the Riemann tensor but the context should always distinguish the two), given by contracting the Ricci tensor with the metric inverse: $R = R^a_{\ a} = g^{ab}R_{ab} = g^{ab}R^c_{\ acb}$.

Symmetry (5) is known as the Bianchi identity. Contracting the Bianchi identity leads to an important result satisfied by the Ricci tensor and the scalar curvature which we now derive. By contracting the indices e and a, we obtain

$$\nabla_a R^a_{\ bcd} + \nabla_d R_{bc} - \nabla_c R_{bd} = 0.$$

Raising the index b and contracting it with d gives

$$\nabla_a R^{ab}{}_{cb} + \nabla_b R^b{}_c - \nabla_c R = 0.$$

By applying symmetries (1) and (3), we can write $\nabla_a R^{ab}{}_{cb} = \nabla^a R_{ac}$. Thus we obtain

$$\nabla^a R_{ac} + \nabla^b R_{bc} - \nabla_c R = 0,$$

or equivalently,

$$\nabla^a \left(R_{ab} - \frac{1}{2} R g_{ab} \right) = 0.$$

The $(0,2)$ tensor $G = Ric - \frac{1}{2}Rg$ with components given by $G_{ab} = R_{ab} - \frac{1}{2}Rg_{ab}$ is known as the ***Einstein tensor***. It will play a fundamnetal role in Einstein's field equations.

4.2 Geodesic Deviation

Let (M, g) be a spacetime. Motivated by the equivalence principle we believe that material particles are following timelike paths which locally maximize their proper time. In section 3.4 we saw that timelike geodesics are precisely these paths. Now let's consider two observers who are initially at the same height above the earth. If we release them from rest, they will fall to the Earth each following some timelike geodesic in spacetime. However, these timelike geodesics are in some sense getting "closer" to each other. This is unlike the case in Minkowski space where we would say that the timelike geodesics the observers follow are staying the same distance apart, i.e. they're parallel. We want to find a way to quantify what we mean by geodesics getting closer and that relationship is

given by the curvature tensor.

As in section 3.4 let N_p be a normal neighborhood about a point $p \in M$. Define the two surface $\alpha(s,t) = \exp_p\big(su(t)\big)$ where $g\big(u(t),u(t)\big) = -c^2$. If $v = \partial/\partial s\big|_\alpha$ and $w = \partial/\partial t\big|_\alpha$, then from Lemma 3.4, we saw $g(v,w) = 0$. We can think of $\nabla_v w$ as the rate of change along a geodesic of the displacement to an infinitesimally nearby geodesic, so it measures the spread of nearby geodesics. Similarly, we may interpret $a = \nabla_v(\nabla_v w)$ as how fast the nearby geodesics are spreading. Since v and w are coordinate vector fields, we can also write $a = \nabla_v(\nabla_w v)$. Let $\{e_a\}$ be any vector basis with dual one form basis $\{e^a\}$. Then the components of a are

$$\begin{aligned} a^a &= v^c\nabla_c(w^b\nabla_b v^a) \\ &= (v^c\nabla_c w^b)(\nabla_b v^a) + w^b v^c \nabla_c\nabla_b v^a \\ &= (w^c\nabla_c v^b)(\nabla_b v^a) + w^b v^c\nabla_b\nabla_c v^a + w^b v^c(\nabla_c\nabla_b - \nabla_b\nabla_c)v^a \\ &= w^c\nabla_c(v^b\nabla_b v^a)) + R^a{}_{bcd}v^c w^d v^b \\ &= R^a{}_{bcd}v^c w^d v^b. \end{aligned}$$

The last equality follows since v is tangent to a geodesic. This equation is known as the ***geodesic deviation equation*** (or ***Jacobi equation***). It tells us that the rate at which geodesics spread is determined precisely by the curvature tensor. In our example with the two observers above the Earth, we saw that gravity from the Earth is forcing the geodesics of the two observers to get closer and closer. Therefore by the geodesic deviation equation, we conclude that gravitational effects occur when the curvature is nonzero. Our next task is to determine what controls curvature?

4.3 Stress-Energy-Momentum Tensor

In Newton's theory of gravity, the matter density of space ρ is related to the acceleration of test bodies $\vec{a}$ by ***Poisson's equation***: $\Delta\phi = 4\pi G\rho$, where $\vec{a} = -\text{grad}\ \phi$ and Δ is the Laplacian in $\mathbb{R}^3$ and G is Newton's constant. We seek a coordinate independent way of describing matter in a spacetime.

Material particles in a spacetime (M, g) are timelike curves which have an attribute known as ***rest mass*** $m > 0$. Let γ be a timelike curve for a material particle with rest mass m and τ its proper time. If γ is parametrized by τ, then $g(\gamma', \gamma') = -c^2$ and we call γ' the ***four-velocity*** of γ. The ***momentum*** of γ is defined as the vector $p = m\gamma'$. If λ is another timelike curve also parametrized by proper time, then the ***energy of*** γ ***as measured by*** λ is $E = -g(p, \lambda') = -mg(\gamma', \lambda')$. If γ measures its own energy, then this is $E = -mg(\gamma', \gamma') = mc^2$ which is Einstein's famous formula relating energy to rest mass.

To see how this notion of energy generalizes our familiar understanding of energy, consider a material particle $\gamma(s) = (s, vs, 0, 0)$ in Minkowski space $(\mathbb{R}^4, \eta)$ with rest mass $m > 0$ and an observer $\lambda(s) = (s, 0, 0, 0)$. If τ measures the proper time of γ, then the energy of γ as measured by λ is

$$E = -m\eta(\gamma', \lambda') = mc^2 \frac{ds}{d\tau}.$$

and since

$$\tau = \frac{1}{c}\int \sqrt{-\eta\big(\gamma'(s), \gamma'(s)\big)ds} = \frac{1}{c}\int \sqrt{c^2 - v^2}ds = s\sqrt{1 - v^2/c^2},$$

we have

$$E = \frac{mc^2}{\sqrt{1 - v^2/c^2}} = mc^2 \left(1 + \frac{v^2}{2c^2} + \frac{3v^4}{8c^4} + \cdots\right) \approx mc^2 + \frac{1}{2}mv^2.$$

where we recognize $\frac{1}{2}mv^2$ as the kinetic energy of a particle with mass m and speed v. Thus at speeds with low velocity, E represents the usual energy from kinematics. If we want our speed to approach c, then E must approach infinity and note that this crucially relies on the fact that $m > 0$. This is why we define material particles strictly as *timelike* curves. In other words, material objects don't travel faster than the speed of light.

To discuss continuous matter distributions, we need a $(0, 2)$ tensor T called the ***stress-energy-momentum tensor***. For a timelike observer γ, $T(\gamma', \gamma')$ represents the mass-energy per unit volume, as measured by γ. If x is a vector orthogonal to γ', then $T(x, \gamma')$ is interpreted as the momentum density of the matter in the x-direction, and if y is also orthogonal, then $T(x, y)$ is interpreted the ***stress*** of the material objects in the x and y directions. We will abbreviate the name and usually refer to T as the ***stress tensor***.

We require two properties of the stress tensor. (1) T is symmetric. (2) The components of T in any vector basis $\{e_a\}$ satisfy $\nabla_a T^{ab} = 0$. In the presence of a Killing vector field k, these two properties give rise to a conservation law. To see this, define the vector p by $p^a = T^{ab}k_b$. Then

$$\nabla_a p^a = k_b \nabla_a T^{ab} + T^{ab} \nabla_a k_b.$$

The first term is zero by (2) and the second is zero since T^{ab} is symmetric and $\nabla_a k_b + \nabla_b k_a = 0$ since k is a Killing field. Therefore $\nabla_a p^a = 0$. If k is taken as $\partial/\partial t$ in Minkowski

space, then this is the familiar concept of conservation of energy. If $k = \partial/\partial x$, then this is the familiar concept of conservation of momentum in the x-direction. However, notice that in an arbitrary spacetime (M, g), there is no guarantee a Killing field will exist.

The most important example of a continuous matter distribution is that of a ***perfect fluid***. Let u be the unit timelike vector field which represents the four-velocities of the matter. Then a perfect fluid has a stress energy tensor with components

$$T_{ab} = \left(\rho + \frac{P}{c^2}\right) u_a u_b + P g_{ab}.$$

The functions ρ and P are the ***mass-energy density*** and ***pressure density*** of the matter as measured by the matter, respectively. By projecting the conservation equation $\nabla^a T_{ab} = 0$ onto the parallel and perpendicular components to u^a, we get:

$$c^2 u_a \nabla^a \rho - (P + c^2\rho)\nabla^a u_a = 0,$$

$$(P + c^2\rho) u^a \nabla_a u_b + (c^2 g_{ab} + u_a u_b)\nabla^a P = 0.$$

Let us consider these two equations in the nonrelativistic limit: $\nabla^a = \partial^a$ for coordinates (t, x, y, z) in Minkowski space, $\frac{P}{c^2} << \rho$, $u^a = (c, \vec{u})$, and $\frac{|\vec{u}|}{c^2}\frac{dP}{dt} << |\vec{\nabla} P|$. Then these equations produce

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{u}) = 0,$$
$$\rho \left[\frac{\partial \vec{u}}{\partial t} + (\vec{u} \cdot \vec{\nabla})\vec{u}\right] = -\vec{\nabla} P$$

The first equation is the familiar conservation of mass and the second equation is Euler's equation for fluid dyanmics.

4.4 Einstein's Field Equations

In Newtonian mechanics, it is mass which determines the motion of material particles. In chapter 3, we saw that material particles are following geodesics which are determined by the metric on the spacetime manifold. Therefore we seek a relationship which bridges the matter content, i.e. the stress tensor, and the metric. We do this by asking ourselves how two material particles in a gravitational field, e.g. above the Earth's surface, will accelerate towards each other. In Newtonian mechanics the acceleration between the two material particles, which are separated by a vector $\vec{w}$, is given by $-(\vec{w}\cdot\vec{\nabla})\vec{\nabla}\phi$ where ϕ is determined by Poisson's equation $\Delta\phi = 4\pi G\rho$ and Δ is the Laplacian in $\mathbb{R}^3$. However, we saw from the geodesic deviation equation that the rate at which two nearby geodesics spread is given by $R^a{}_{bcd}v^c w^d v^b$ where $v = \partial/\partial s\big|_\alpha$ and $w = \partial/\partial t\big|_\alpha$ are orthogonal vector fields defined on the two-surface $\alpha(s,t) = \exp_p\big(su(t)\big)$ where $g\big(u(t),u(t)\big) = -1$. If the two geodesics are coming closer together, then we expect that the two particles represented by the geodesics are being pulled by some gravitational force. This suggests a correspondence between the two terms

$$R^a{}_{bca}v^c v^b \text{ and } -\Delta\phi.$$

But $\Delta\phi = 4\pi G\rho$ and we know that $T_{ab}v^a v^b = \rho$ if $P = 0$. Thus suggests that the curvature and the stress-energy tensor are related by the following equation

$$R^a{}_{bca}v^c v^b = -\frac{4\pi G}{c^4}T_{cd}v^c v^d$$

The factor c^{-4} is necessary to get the correct units. Since the curvature tensor satisfies the symmetry $R^a{}_{bca} = R^a{}_{bac} = R_{bc}$, this suggests the following equation

$$R_{bc} = \frac{4\pi G}{c^4} T_{bc}.$$

Indeed this equation was postulated by Einstein, but it leads to unphysical constraints on the universe. To see this, recall that we wanted our stress tensor to satisfy $\nabla^b T_{bc} = 0$. Then the above equation would imply $\nabla^b R_{bc} = 0$. But by the Einstein tensor, we would have

$$0 = \nabla^b G_{bc} = \nabla^b \left(R_{bc} - \frac{1}{2} R g_{bc} \right) = -\frac{1}{2} g_{bc} \nabla^b R$$

$\nabla^b R = 0$ implies R is constant through the universe. Hence $T = T^a_a$ is constant throughout the universe. This constraint is highly unphysical and unmotivated, so we disregard the relation $R_{bc} = \frac{4\pi G}{c^4} T_{bc}$ and seek a better one. Both the Einstein tensor G_{ab} and the stress-energy tensor T_{ab} vanish when they're covariantly differentiated (i.e. $\nabla^a(T_{ab}) = \nabla^a G_{ab}) = 0$). This suggests the following relation

$$G_{ab} = R_{ab} - \frac{1}{2} R g_{ab} = \frac{8\pi G}{c^4} T_{ab}.$$

These are ***Einstein's field equations***. By taking the trace of the above equation, we see that $R = -\frac{8\pi G}{c^4} T$, so we can rewrite the equations as

$$R_{ab} = \frac{8\pi G}{c^4} \left(T_{ab} - \frac{1}{2} g_{ab} T \right).$$

Therefore by imposing realistic energy conditions (i.e. restricting certain values of T_{ab}), we can control the Ricci curvature. This idea plays a fundamental role in the singularity theorems. Also, notice that when $T \approx \rho$ (e.g. in the Newtonian limit), we recover our

first assumption $R_{bc} \approx \frac{4\pi G}{c^4} T_{bc}$.

Notice that our construction of the Einstein's field equations was not unique. For any number Λ, we can define a new Einstein tensor $\widetilde{G}_{ab} = G_{ab} + \Lambda g_{ab}$. This new Einstein tensor will satisfy $\nabla^a \widetilde{G}_{ab} = 0$, so one can postulate $\widetilde{G}_{ab} = \frac{4\pi G}{c^4} T_{ab}$ as the Einstein's field equations. In this case we call Λ a ***cosmological constant***. Alternatively, one can define $\widetilde{T}_{ab} = T_{ab} - \frac{c^4}{4\pi G} \Lambda g_{ab}$ so that the Einstein's field equations look like $G_{ab} = \frac{4\pi G}{c^4} \widetilde{T}_{ab}$. In this case, Λ is referred to as ***dark energy***. Introducing the term Λ doesn't change the complexity of Einstein's equations, and so for most of this thesis we will disregard it. Or we will just assume it's incorporated in the stress-tensor as dark energy.

When Λ is thought of a cosmological constant, current observations put Λ at a small but nonzero positive quantity.

5 The Schwarzschild and Friedman-Robertson-Walker Solutions

A ***solution*** of Einstein's equations is a spacetime (M, g) for which the Einstein field equations are satisfied for some stress tensor T. Because the field equations are so complex, we can only hope to find solutions with a high degree of symmetry. For example, in section 5.1 we will describe the Schwarzschild solution which describes spacetime outside of a star. Since stars are observed to be spherical we will assume that M possesses some spherical symmetry. In Section 5.2 we will find the FRW solution which describes the

whole universe. Since the universe seems to look the same in every direction, we will assume M possesses isotropic properties. These solutions with a high degree of symmetry are only idealized models of what we believe the actual spacetime to be, nevertheless, they give us means of experimentally testing Einstein's theory. Moreover, they give us hints of pathological global behavior. In the Schwarzschild solution, observers can end their existence in a finite amount of proper time. Likewise, in the FRW solution, every observer begins their existence in a finite amount of proper time. It was once thought that this pathological behavior was a result of the high degree of symmetry in these solutions, however the singularity theorems will show that this pathological behavior exists in spacetimes without symmetry.

5.1 The Schwarzschild Solution

We are interested in solving Einstein's equations for the gravitational field outside a stellar object (e.g. the Sun, the Earth, etc.) Since large stellar objects are nearly spherical, we will assume that a spacetime (M, g) is ***spherical symmetric***. Physically, this means that g is invariant under rotations which is what we expect from the gravitational field outside the sun. Mathematically, this means that the isometry group of (M, g) contains a subgroup isomorphic to SO(3) and the orbits of this subgroup are two-dimensional spheres. Moreover, we also assume (M, g) is ***static***. Physically, this means that the Sun's gravitational field "doesn't change with time." Mathematically, this means that there exists a unique one-parameter group of isometries, $\{\phi_t\}$, whose orbits are timelike

curves and there exists a foliation of spacelike hypersurfaces, $\{\Sigma_t\}$, which are everywhere orthogonal to the orbits. The one-parameter group of isometries $\{\phi_t\}$ generate a timelike Killing vector field k_t which is tangent to the orbits of ϕ_t. These assumptions allow us to introduce convenient coordinates on M. First introduce coordinates $\{x^1(t), x^2(t), x^3(t)\}$ on Σ_t. Let $h(t)$ be the induced metric on Σ_t with components, $h_{ab}(t)$ $(a, b = 1, 2, 3)$, in terms of the given coordinates coordinates, so as long as $k_t \neq 0$ on Σ_t, $\{ct, x^1(t), x^2(t), x^3(t)\}$ are coordinates for M. Since k_t is a Killing vector field, g must be independent of t, so the orthogonality condition allows us to write the metric as

$$g = -|g(k_t, k_t)|(cdt)^2 + \sum_{a,b=1}^{3} h_{ab} dx^a dx^b.$$

Now by spherical symmetry, g induces a metric on each orbit two-sphere which, by the symmetry, must be a positive multiple of the metric on a unit two-sphere: $r^2(d\theta^2 + \sin^2\theta d\phi^2)$ where $r = \sqrt{A/4\pi}$ and A is the area of the two-sphere. Now the spherical symmetry and uniqueness of k_t imply that k_t is orthogonal to all all the orbit two-spheres. Therefore each two-sphere must lie within some spacelike Σ_t, so as long as $\nabla r \neq 0$, (r, θ, ϕ) are coordinates for Σ_t. In fact, the metric $h(t)$ on Σ_t must only depend on r by spherical symmetry. Thus the spacetime metric in the coordinates (ct, r, θ, ϕ) takes the form

$$g = -\alpha(r)(cdt)^2 + \beta(r)dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)$$

where $\alpha(r)$ and $\beta(r)$ are positive functions of r. It should be pointed out that these coordinates are only valid when $k_t \neq 0$ and $\nabla r \neq 0$. This will become important when discussing singularities of the Schwarzschild metric.

We're mainly concerned with solutions that exist outside the Sun. In this region,

there is no matter or energy so we assume the stress tensor satisfies $T = 0$. Therefore by Einstein's field equations, we have $Ric = 0$. So determine $\alpha(r)$ and $\beta(r)$, we solve

$$\begin{aligned} 0 &= R_{ab} \\ &= R^{c}{}_{acb} \\ &= \frac{\partial \Gamma^{c}{}_{ba}}{\partial x^{c}} - \frac{\partial \Gamma^{c}{}_{ca}}{\partial x^{b}} + \Gamma^{c}{}_{cd}\Gamma^{d}{}_{ba} - \Gamma^{c}{}_{bd}\Gamma^{d}{}_{ca} \end{aligned}$$

and

$$\Gamma^{a}{}_{bc} = \frac{1}{2} g^{ad} \left(\frac{\partial g_{cd}}{\partial x^{b}} + \frac{\partial g_{bd}}{\partial x^{c}} - \frac{\partial g_{bc}}{\partial x^{d}} \right).$$

Working through all the components, we find

$$\alpha(r) = 1 + \frac{C}{r} \quad \text{and} \quad \beta(r) = \left(1 + \frac{C}{r}\right)^{-1},$$

where C is an undetermined constant, so C is a parameter to the set of solutions of Einstein's equations which are static and spherically symmetric.

In fact, there is a good choice to choose for C which is related to the mass of the stellar object. To find this, let us consider the timelike geodesics in M. Recall that these are the paths followed by observers in M, so timelike geodesics can be considered the paths of planets around the Sun given that the planets own mass don't add any significant contributions to the stress tensor. Let γ be a timelike geodesic which is parametrized by its proper time τ. In our coordinates, we can write $\gamma(\tau) = \big(t(\tau), r(\tau), \theta(\tau), \phi(\tau)\big)$. Then we have

$$\begin{aligned} -c^2 &= g(\gamma', \gamma') \\ &= g_{ab} \frac{dx^a}{d\tau} \frac{dx^b}{d\tau} \\ &= -c^2 \left(1 - \frac{C}{r}\right) \left(\frac{dt}{d\tau}\right)^2 + \left(1 - \frac{C}{r}\right)^{-1} \left(\frac{dr}{d\tau}\right)^2 + r^2 \left(\frac{d\phi}{d\tau}\right)^2 + r^2 \sin^2\theta \left(\frac{d\theta}{d\tau}\right)^2 \end{aligned}$$

Notice that the metric is independent of $\partial/\partial t$ and $\partial/\partial\phi$ so each are Killing fields. Recall that for any geodesic γ and any killing field k, the quantity $g(\gamma', k)$ is conserved along γ. Therefore we have two conserved quantities on γ

$$E = -g\left(\gamma', \frac{\partial}{\partial t}\right) = -c^2\left(1 - \frac{C}{r}\right)\frac{dt}{d\tau}$$
$$L = g\left(\gamma', \frac{\partial}{\partial \phi}\right) = r^2 \sin^2\theta \frac{d\phi}{d\tau}.$$

The conserved quantity L allows us to restrict motion of γ to within "the plane" $\theta = \pi/2$. To see this, pick any time τ. We can find an isometry of the metric such that $\phi(\tau) = 0$ and $d\phi/d\tau\big|_\tau = 0$ under this isometry. This implies L vanishes all along γ which means that $d\phi/d\tau = 0$ along γ. Therefore γ is restricted to "the plane" $\phi = 0$. Now choose an isometry that maps the plane $\phi = 0$ to the plane $\theta = \pi/2$. Thus $\theta = \pi/2$ along γ. So now we have

$$-c^2 = -c^2\left(1 - \frac{C}{r}\right)\left(\frac{dt}{d\tau}\right)^2 + \left(1 - \frac{C}{r}\right)^{-1}\left(\frac{dr}{d\tau}\right)^2 + r^2\left(\frac{d\phi}{d\tau}\right)^2 + r^2\sin^2\theta\left(\frac{d\theta}{d\tau}\right)^2$$
$$-c^2\left(1 - \frac{C}{r}\right) = -\frac{E^2}{c^2} + \left(\frac{dr}{d\tau}\right)^2 + \left(1 - \frac{C}{r}\right)\frac{L^2}{r^2}.$$

Rearranging the above equation, we get

$$\frac{1}{2}\left(\frac{dr}{d\tau}\right)^2 + \left(-\frac{c^2 C}{2r} + \frac{L^2}{2r^2} - \frac{CL^2}{2r^3}\right) = \frac{\frac{E^2}{c^2} - c^2}{2}.$$

If we let $U(r) = -\frac{c^2 C}{2r} + \frac{L^2}{2r^2} - \frac{CL^2}{2r^3}$, the above equation resembles conservation of energy

$$\text{Kinetic Energy} \;+\; \text{Potential Energy} \;=\; \text{Total Energy}.$$

In fact, if we consider the Newtonian limit $\tau \approx t$ and let $C = 2Gm/(c^2 r)$ (where m is the mass of the stellar object), then we recover Newton's law of planetary motion with

L acting as the angular momentum of the particle represented by the geodesic γ:

$$\frac{1}{2}\left(\frac{dr}{dt}\right)^2 + \left(-\frac{Gm}{r} + \frac{L^2}{r^2} - \frac{GmL^2}{c^2r^3}\right) = \text{ Total Energy.}$$

We are able to reproduce Newtonian Laws from general relativity which gives credence to the theory. The only unfamiliar term above is $GmL^2/(c^2r^3)$, but this term is negligble for everyday objects like the planets and asteroids orbiting the Sun. However it should be noted that it is precisely this term that predicts the discrepancy of Mercury's orbit from the classical Newtonian limit. This discrepancy was a problem for physicists in the 19th century leading to predictions of unobserved planets. The fact that general relativity can explain this discrepancy is regarded as one of its main successes.

Now we're able to define the ***Schwarzschild solution***. It's a solution (M, g) of Einstein's equations *outside* a spherical object with mass m and can be given ***Schwarzschild coordinates*** (ct, r, θ, ϕ) such that

$$g = -\left(1 - \frac{2Gm}{c^2r}\right)(cdt)^2 + \left(1 - \frac{2Gm}{c^2r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2).$$

These coordinates cannot be extended to $\theta = 0, \pi$, $r = 2Gm/c^2$, or $r = 0$. Now $\theta = 0, \pi$ are examples of ***coordinate singularities***, points where the metric would be degenerate but a change of coordinates removes the degeneracy. Nothing bad is going on at coordinate singularities, just a poor choice of coordinates.

Therefore we really only need to worry about the $r = 2Gm/c^2$ and $r = 0$. In fact these may not even pose a problem. If the stellar object has a radius which is larger than $r = 2Gm/c^2$, then Schwarzschild coordinates aren't appropriate to describe $r \leq 2Gm/c^2$ since the coordinates are only valid outside the star, i.e. where stress tensor is zero. However, astronomical predictions of spherical stellar objects suggest that if the stellar

object has a mass greater than 1.5 times that of our Sun's, then the stellar object will run of its nuclear fuel to keep it from collapsing in on itself, in which case it will the matter will just keep collapsing forever. Although this type of situation is not static, it tells us that the region $0 < r < 2Gm/c^2$ is physically relevant and so we must concern ourselves with $r = 2Gm/c^2$ and $r = 0$.

We will show $r = 2Gm/c^2$ is merely a coordinate singularity. Define $r^*(r) = r + \frac{2Gm}{c^2} \log\left(\frac{rc^2}{2Gm} - 1\right)$ and $v(t, r) = t + r^*(r)$, then we have the ***Eddington-Finkelstein coordinates***, (v, r, θ, ϕ). In terms of these coordinates, the metric is given by

$$g = -\left(1 - \frac{2Gm}{c^2 r}\right) dv^2 + (dvdr + drdv) + r^2(d\theta^2 + \sin^2\theta d\phi^2).$$

Although $g_{vv} = 0$ at $r = 2Gm/c^2$, the metric is still nondegenerate since its determinant in these coordinates is $-r^4 \sin^2\theta \neq 0$. Therefore $r = 2Gm/c^2$ is only a coordinate singularity.

Now we can ask if $r = 0$ is also a coordinate singularity. For example $r = 0$ is a coordinate singularity for $\mathbb{R}^2$ with polar coordinates (r, θ) and metric $dr^2 + r^2 d\theta^2$. One way to see if $r = 0$ is *not* a coordinate singularity is if we can find a coordinate-independent quantity that behaves poorly at $r = 0$. The easiest such quantity to consider is a scalar derived from the curvature tensor. If such a scalar diverges at the coordinate point of interest, then the ccoordinate point is called a ***curvature singularity***. For Schwarzschild coordinates, one can show that $R^{abcd}R_{abcd} = 48G^2m^2/(c^4r^6)$. Thus $r = 0$ is a curvature singularity.

Thus we can conclude that the manifold M can not be extended to $r = 0$. The manifold is breaking at these points because the curvature is blowing up there. Now an

important question to ask is how long does it take observers to reach the point $r = 0$? For example, if it takes observers an infinite amount of proper time to reach the point $r = 0$, then this curvature singularity may not be physically relevant since "it takes an infinite amount of time to get there." We will show that this is not the case. Observers following timelike geodesics can reach the coordinate $r = 0$ in a finite amount of proper time.

Let's imagine an observer following a timelike geodeisc, $\gamma(\tau)$, who has unfortunately found him or herself in the region $r < 2Gm/c^2$. In this region we have $\frac{L^2}{r^2} - \frac{GmL^2}{2c^2r^3} < 0$. Therefore

$$\left(\frac{dr}{d\tau}\right)^2 \geq \frac{E^2}{c^2} - c^2.$$

Moreover $\partial/\partial r$ is timelike in this region so r either increases with τ or decreases with τ. Therefore if we assume the observer initially starts with $dr/d\tau < 0$, then $dr/d\tau < 0$ all along γ. Thus we can integrate the above inequality from any $r_0 < 2Gm/c^2$ to $r = 0$ and find that the total elapsed proper time for this path satisfies

$$\tau \leq \frac{r_0}{\sqrt{E^2/c^2 - c^2}}.$$

This shows something catastrophic for the observer γ. In a finite amount of time he reaches the curvature singularity and after that he ceases to exist.

5.2 Friedman-Robertson-Walker Solutions

In this section we want to find a solution, (M, g), of Einstein's equations of the entire universe. This is a daunting task since we would have to know complete knowledge of the

stress tensor at every point of our spacetime manifold. The task can be greatly simplified if we can make some assumptions about our universe.

Since the time of Copernicus, it has been believed that we do not occupy a special region in the universe. We are merely on an average planet, orbiting an average star, within an average galaxy, which is itself within an average cluster of galaxies. If we're not special, then we shouldn't expect anyone else to be special. Therefore it is believed that our spacetime satisfies a homogeneity property - the characteristics of our surroundings would appear the same no matter where we are in M. Similarly, any direction we look out at in space appears no different than any other direction. We see roughly the same distribution of galaxies and galaxy clusters no matter where we look. This condition is known as isotropy. Precise mathematical definitions of homogeneity and isotropy for a spacetime are given below, but first let's describe an analogy which helps clarify the concepts. Imagine yourself as an ant in a sandbox. No matter where the ant is, the sandbox looks roughly the same. Moreover, it doesn't matter if he looks north, south, east, or west; each direction looks roughly the same. This is isotropy. If someone were to rake the sandbox uniformly in one direction, then the sandbox would still be homogeneous, but it would no longer be isotropic.

A spacetime (M, g) is said to be ***spatially homogeneous*** if there exists a one-parameter family of spacelike hypersurfaces, $\{\Sigma_t\}$, foliating M such that for each t and any points $p, q \in \Sigma_t$, there exists an isometry ϕ_t of (M, g) which takes p into q. (M, g) is said to be ***spatially isotropic*** if there exists a congruence of timelike curves $\{\gamma\}$ such that at any point $p \in M$ and spacelike vectors $u, v \in T_pM$, there is an isometry ψ of (M, g) which leaves p and $\gamma'|_p$ fixed but maps u to v. The curves $\{\gamma\}$ would represent

the worldlines of galaxies and will be called ***isotropic observers***.

If (M, g) is spatially homogeneous and spatially isotropic, then the congruence of timelike curves, $\{\gamma\}$, must be orthogonal to the one-parameter of spacelike hypersurfaces, $\{\Sigma_t\}$. Moreover, by homogeneity the isotropic observers $\{\gamma\}$ must agree on the time difference between any two hypersurfaces Σ_t and $\Sigma_{t'}$. Therefore if $h(t)$ is the Riemannian metric induced from g on Σ_t, then we can write

$$g = -c^2 d\tau^2 + h(\tau)$$

where τ is the proper time as measured by the isotropic observers $\{\gamma\}$ and $h(\tau)$ is really $h\big(t(\tau)\big)$. If Ric_h is the Ricci curvature for $\big(\Sigma_t, h(t)\big)$, then the isotropic condition implies that Ric_h is a multiple of g and for $n = 3$, $\big(\Sigma_t, h(t)\big)$ is a space of constant sectional curvature. It is a standard result from Riemannian geometry that the spaces of constant curvature are locally isometric to the sphere, Euclidean space, and hyperbolic space. We can locally cover these spaces with coordinates (r, θ, ϕ) so that g has the following form

$$g = -c^2 d\tau^2 + a^2(\tau)\left[\frac{dr^2}{1 - kr^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2)\right],$$

where $k = +1, 0, -1$ corresponds to the sphere, Euclidean space, and hyperbolic space, respectively. These solutions are known as the ***Friedman-Robertson-Walker (FRW) solutions***. $a(\tau)$ is known as the ***scale factor*** and it determines the spatial expansion of the spacelike hypersurfaces $\{\Sigma_t\}$.

Notice that if D is the Riemannian distance between two points on Σ_t, then D is proportional to a. Therefore

$$\frac{dD}{d\tau} = \frac{D}{a}\frac{da}{d\tau} = HD$$

where the function $H = \frac{1}{a}\frac{da}{d\tau}$ is known as ***Hubble's constant*** even though it is technically not a constant. The "linear" relation $dD/d\tau = HD$ can be experimentally verified by measuring the Doppler shift of distant galaxies. These measurements were in fact made by Hubble which gave credence to the FRW solutions.

Since our spacetime is suppose to model the observable universe, we can assume each galaxy is like a "grain of dust." Let $u = \partial/\partial\tau$ be the four-velocities of the isotropic observers. Then a stress-energy tensor which adequately models "dust" is $T_{ab} = \rho u_a u_b$. Moreover, measurements of the cosmic microwave background show that there is a thermal distribution of radiation pressure at a temperature of 3 Kelvin which fills the universe. For these reasons, we assume the stress energy tensor takes the form of a perfect fluid

$$T_{ab} = \left(\rho + \frac{P}{c^2}\right) u_a u_b + P g_{ab}.$$

Now we set out to solve Einstein's equations,

$$G_{ab} = R_{ab} - \frac{1}{2} R g_{ab} = \frac{8\pi G}{c^4} T_{ab},$$

in hopes of finding an equation that describes how $a(\tau)$ evolves. The first step is to solve for the Ricci tensor components and the scalar curvature. Using the coordinates (τ, r, θ, ϕ), we calculate the Christoffel symbols using the formula

$$\Gamma^c{}_{ab} = \frac{g^{cd}}{2}\left(\frac{\partial g_{bd}}{\partial x^a} + \frac{\partial g_{ad}}{\partial x^b} - \frac{\partial g_{ab}}{\partial x^d}\right).$$

The nonzero components are

$$\Gamma^\tau{}_{rr} = c^{-2}\frac{a\dot{a}}{1-kr^2}, \quad \Gamma^\tau{}_{\theta\theta} = c^{-2} r^2 a\dot{a}, \quad \Gamma^\tau{}_{\phi\phi} = c^{-2} r^2 \sin^2\theta a\dot{a},$$

$$\Gamma^r{}_{r\tau} = \Gamma^\theta{}_{\theta\tau} = \Gamma^\phi{}_{\phi\tau} = \frac{\dot{a}}{a},$$

$$\Gamma^{r}{}_{rr} = \frac{rk}{1-kr^2}, \quad \Gamma^{r}{}_{\theta\theta} = r(kr^2-1), \quad \Gamma^{r}{}_{\phi\phi} = r\sin^2\theta(kr^2-1),$$

$$\Gamma^{\theta}{}_{\theta r} = \frac{1}{r}, \quad \Gamma^{\theta}{}_{\phi\phi} = -\cos\theta\sin\theta,$$

$$\Gamma^{\phi}{}_{\phi r} = \frac{1}{r}, \quad \Gamma^{\phi}{}_{\phi\theta} = \cot\theta.$$

where $\dot{a} = da/d\tau$.

From here we calculate the Ricci tensor components by the formula

$$R_{ab} = R^{c}{}_{acb} = \frac{\partial\Gamma^{c}{}_{ba}}{\partial x^c} - \frac{\partial\Gamma^{c}{}_{ca}}{\partial x^b} + \Gamma^{c}{}_{cd}\Gamma^{d}{}_{ba} - \Gamma^{c}{}_{bd}\Gamma^{d}{}_{ca}.$$

First, we find

$$R_{\tau\tau} = -\frac{3\ddot{a}}{a}.$$

Now we use the symmetries in the metric to help simplify the problem. Since the manifold is spatially isotropic, if s is any unit spacelike vector which is orthogonal to the isotropic observers u, then the quantity $R_{ab}s^a s^b$ doesn't depend on choice of s (if it did one can show that spatial isotropy is violated). This implies

$$R_{ab}s^a s^b = \frac{1-kr^2}{a^2}R_{rr} = \frac{1}{r^2a^2}R_{\theta\theta} = \frac{1}{r^2a^2\sin^2\theta}R_{\phi\phi}.$$

From direct calculation, we find

$$R_{\theta\theta} = \frac{r^2}{c^2}(\ddot{a}a + 2\dot{a}^2) + 2kr^2.$$

Therefore

$$R_{ab}s^a s^b = \frac{1}{c^2}\left(\frac{\ddot{a}}{a} + 2\frac{\dot{a}^2}{a^2}\right) + 2\frac{k}{a^2}.$$

The scalar curvature is then

$$R = g^{ab}R_{ab} = -c^{-2}R_{\tau\tau} + 3R_{ab}s^a s^b = \frac{6}{c^2}\left(\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2}\right) + 6\frac{k}{a^2}.$$

Thus Einstein's equations give

$$8\pi G\rho = \frac{8\pi G}{c^4}T_{\tau\tau} = G_{\tau\tau} = R_{\tau\tau} + \frac{c^2}{2}R = 3\frac{\dot{a}^2}{a^2} + 3\frac{c^2k}{a^2}$$

$$\frac{8\pi G}{c^4}P = \frac{8\pi G}{c^4}T_{ab}s^a s^b = G_{ab}s^a s^b = R_{ab}s^a s^b - \frac{1}{2}R = -\frac{1}{c^2}\left(2\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2}\right) - \frac{k}{a^2}.$$

Using the first equation, we can rewrite the second equation as

$$3\frac{\ddot{a}}{a} = -\frac{4\pi G}{c^4}\left(\rho + 3\frac{P}{c^2}\right).$$

This equation along with

$$3\frac{\dot{a}^2}{a^2} = 8\pi G\rho - 3\frac{c^2k}{a^2}$$

are known as the ***Friedmann Equations***; they describe the evolution of the scale factor $a(\tau)$.

If we make the physically reasonable assumption that $\rho + 3P/c^2 > 0$, then the first Friedmann equation implies $\ddot{a} \neq 0$, therefore $\dot{a} \neq 0$ almost everywhere. Observations made by Hubble imply that the universe is currently expanding, $\dot{a} > 0$. Since $\ddot{a} < 0$, the universe must have been expanding at a faster rate as one goes backwards in proper time τ of the isotropic observers. The two conditions $\dot{a} > 0$ now and $\ddot{a} < 0$ always imply that $a(\tau)$ must cross the τ axis at which point $a = 0$ and the metric become singular. Thus the isotropic observers find themselves existing for only a finite amount of time in the past. Notice that since $R = \frac{6}{c^2}\left(\frac{\ddot{a}}{a} + 2\frac{\dot{a}^2}{a^2}\right) + 2\frac{k}{a^2}$, we have a curvature singularity at the proper time when $a = 0$.

6 Causal Structure

Except for the exact solutions of Einstein's equations we found in chapter 5, our results regarding a spacetime have only been local. If we want to understand global phenomenon, such as the birth or end of our universe, then we want to understand how the spacetime manifold behaves globally; this global behavior is commonly referred to as ***causal structure***. In this chapter we lay out the basic definitions and results of causal structure. The results developed here are both interesting on their own and are essential to understanding the singularity theorems in chapter 7.

Recall that if (M, g) is a space-time, the set of null vectors in T_pM is defined as the lightcone. The lightcone minus the null vector is a topological space with exactly two disconnected components, we arbitrarily call one-component ***future directed*** and the other ***past directed***. Timelike (null) vectors that point within (on) the futured directed component are also called future directed. A curve γ is a ***future directed timelike curve*** if γ' is a future directed timelike curve everywhere along γ. Likewise, γ is a ***future directed causal curve*** if γ' is either a future directed timelike or null (but nonzero) vector everywhere along γ. Analogous definitions apply to past directed vectors and curves. If we can make a continuous choice of future and past as p varies in M then we call (M, g) ***time-orientable***. There are multiple ways to make this notion precise, but they are all equivalent to the following fact: *A spacetime is time-orientable if and only if there exists a smooth nonvanishing timelike vector field v on M*. The three examples of spacetimes that we have seen: Minkowski space, the Schwarzschild solution, and the FRW solutions are all time-orientable. Nonetheless, it is easy to construct ex-

amples of non-simply connected spacetimes that are not time orientable. A spacetime that is not time-orientable, has the pathological property that we can not discern the future from the past which is intuitively absurd. However, our intuition is based on the experience from a small portion of the spacetime manifold. Nonetheless, we make the somewhat reasonable assumption that our space-time is time-orientable because it would be almost impossible to find results in causal structure otherwise.

6.1 Future and Past sets

For the rest of this thesis, our space-times (M, g) will assumed to be time-orientable unless otherwise stated. The ***timelike future*** of $p \in M$, denoted by $I^+(p)$, is defined to be the set of points $q \in M$ such that there exists a future directed timelike curve γ which begins at p and ends at q. The ***causal future*** of $p \in M$, denoted by $J^+(p)$, is defined to be the set of points $p \in M$ such that there exists a future directed causal curve γ beginning at p and ending at q. The ***timelike past*** and ***causal past*** of $p \in M$, denoted by $I^-(p)$ and $J^-(p)$, respectively, are defined similarly but with "future" replaced by "past." Also, for any subset $S \subset M$ we put $I^{\pm}(S) = \bigcup_{p \in S} I^{\pm}(p)$ and $J^{\pm}(S) = \bigcup_{p \in S} J^{\pm}(p)$. We will derive results for "+" sets but analogous results hold for "-" sets, simply by redefining which is "future" and which is "past." The sets $J^+(p)$ differ from $I^+(p)$ in the following fundamental way:

Proposition 6.1 *If $q \in J^+(p) \setminus I^+(p)$, then any future directed causal curve connecting p to q is a null geodesic.*

Proof. Let γ be a future directed causal curve connecting p to q which is not a null geodesic. Define a convex normal neighborhood at each point of γ. Since the image of γ is compact, extract a finite number of such neighborhoods which cover γ. Call these sets $\{U_1, \ldots, U_k\}$. γ fails to be a null geodesic in one of these neighborhoods, let's say U_i, then by proposition 3.5 we could deform γ into a timelike curve within U_i. Now γ fails to be a null geodesic in U_{i-1} and U_{i+1}, so we can deform γ into a timelike curve in these neighborhoods as well. Continuing this process through all the neighborhoods, we can find a timelike curve λ (which is the deformed curve of γ) that connects p to q. Hence $q \in I^+(p)$. □

Here are some topological facts:

Proposition 6.2 *Let $S \subset M$.*

(a) $I^+(S)$ *is open.*

(b) $\mathrm{int}\big(J^+(S)\big) = I^+(S)$.

(c) $J^+(S) \subset \overline{I^+(S)}$.

(d) $\partial I^+(S) = \partial J^+(S)$.

Proof. (a) It suffices to show $I^+(p)$ is open for $p \in M$. Fix $q \in I^+(p)$ and let γ be the future directed timelike curve from p to q. Let U be a normal neighborhood about $q \in m$ and let $r \in U$ be a point such that r lies on γ. The preimage of $I^+(r) \cap U$ under $\exp_q$ is an open set, hence itself is an open set contained in $I^+(p)$.

(b) Fix $p \in \mathrm{int}\big(J^+(S)\big)$. There is a convex normal neighborhood U around p which

is contained in $J^+(S)$. Therefore we can construct a future directed causal curve γ which starts in S, ends at p, and is timelike within U. γ is not a null geodesic so by proposition 6.1, $p \in I^+(S)$. Now suppose $p \in I^+(S)$. $I^+(S) \subset J^+(S)$ is an open set, so $p \in \text{int}J^+(S)$.

(c) Fix $p \in J^+(S)$. Let U be any open neighborhood of p and γ a future directed causal curve connecting S to p. Let $r \in I^+(p) \cap U$ and λ the timelike curve which connects p to r. Then $\gamma \cup \lambda$ is a causal curve which connects S to r but is not a null geodesic. Therefore $r \in I^+(S)$. Hence $p \in \overline{I^+(S)}$.

(d) We have $\partial I^+(S) = \overline{I^+(S)} - I^+(S) = \overline{J^+(S)} - \text{int}\big(J^+(S)\big) = \partial J^+(S)$. □

In general $J^{\pm}(p)$ may neither be open nor closed. This can be seen by cutting out points in Minkowski space. Notice that the sets $I^+(S)$ and $J^+(S)$ satisfy the following property: $I^+\big(I^+(S)\big) \subset I^+(S)$ and $I^+\big(J^+(S)\big) \subset J^+(S)$. In general a subset $F \subset M$ which satisfies $I^+(F) \subset F$, then F is called a ***future set***. Similarly, $P \subset M$ is a past set if $I^-(P) \subset P$. We call a subset S ***achronal*** if there exist no two points $p, q \in S$ such that $q \in I^+(p)$. This is equivalent to $I^+(S) \cap S = \emptyset$.

Proposition 6.3 *If F is a future set, then ∂F is achronal.*

Proof. Suppose otherwise. Then there exist points $p, q \in \partial F$ such that $q \in I^+(p)$. Then $p \in I^-(q)$ which is an open set so there exists a neighborhood U about p completely contained in $I^-(q)$. Since $p \in \partial F$, U must intersect F. Therefore there exists a point $r \in U \cap F$, so $p \in I^+(r)$. Therefore $p \in I^+(F)$. Since $I^+(F)$ is open, there exists a neighborhood V around p which is completely contained in $I^+(F)$ but since F is a future set, V is completely contained in F, but this contradicts the initial assumption that

$p \in \partial F$. $\square$

For any achronal set S, we define ***edge***(S) as the set of points $p \in \overline{S}$ such that every neighborhood U of p contains a timelike curve from $I^-(p) \cap U$ to $I^+(p) \cap U$ which does not intersect S. The following theorem shows that achronal sets with no edge points are hypersurfaces in M.

Theorem 6.4 *Suppose S is an achronal set, then S is a three-dimensional C^0 submanifold of M if and only if $S \cap \text{edge}(S) = \emptyset$.*

Proof. First suppose S is a three-dimensional C^0 submanifold of M. Fix $p \in S$ and let U be a connected normal neighborhood about p such that $U \cap S$ is homeomorphic to an open set of $\mathbb{R}^3$. By shrinking U, we may assume $U \setminus S$ has two connected components. Since S is achronal, the open sets $I^{\pm}(p) \cap U$ are open connected sets that are disjoint and do not meet S. Since any future directed timelike curve through p connects $I^-(p) \cap U$ to $I^+(p) \cap U$, $I^-(p) \cap U$ and $I^+(p) \cap U$ are in distinct connected components of $U - S$. But this implies that any timelike curve γ from $I^-(p) \cap U$ to $I^+(p) \cap U$ meets both components. Therefore $\gamma \cap (U \setminus S)$ has two components but $\gamma \cap U$ has one component which implies γ must intersect S. Hence $p \notin \text{edge}(S)$.

Now suppose $S \cap \text{edge}(S) = \emptyset$.Fix $p \in S$. Since S doesn't contain any edge points, we can find a coordinate system $\xi : U \subset M \to \mathbb{R}^4$ such that every timelike curve from $I^-(p) \cap U$ to $I^+(p) \cap U$ intersects A. Choose coordinates (ct, x, y, z) such that $\partial/\partial(ct)$ is future-directed and timelike. We can find a normal neighborhood $V \subset U$ of p such that: 1) $\xi(V) = (a - \delta, a + \delta) \times N \subset \mathbb{R}^1 \times \mathbb{R}^3$ for some $\delta > 0$ and open $N \subset \mathbb{R}^3$.

2) The slices $ct = a$ and $ct = b$ in V are contained in $I^-(p) \cap U$ and $I^+(p) \cap U$, respectively.

For $\vec{x} \in N$, the curve $\xi^{-1}(s, \vec{x})$ for $a \leq s \leq b$ must meet S exactly once. Therefore we have a function $h : N \to (a, b)$ such that $h(\vec{x})$ is the time coordinate of the point where the curve $\xi^{-1}(s, \vec{x})$ meets S. Now let us define the map

$$\phi : S \cap V \to \{t = 0 \text{ slice of } \xi(V)\}$$

by $\phi(q) = \big(ct(q) - h\big(\vec{x}(q)\big), \vec{x}(q)\big)$. Notice that ϕ maps open sets to open sets, so it suffices to show ϕ is continuous, so we need to show h is continuous. Let $\{\vec{x}_n\}$ be a sequence that converges to $\vec{x}$ in N. Assume $\{h(\vec{x}_n)\}$ does not converge to $h(\vec{x})$. Since $\{h(\vec{x}_n)\}$ is a bounded sequence, there is a subsequence $\{h(\vec{x}_{n'})\}$ which converges to some number $d \neq h(\vec{x})$. Let $q = \xi^{-1}\big(h(\vec{x}), \vec{x}\big)$. Notice that $q \in S$ by definition of h and so $\xi^{-1}(d, \vec{x}) \in \big(I^-(q) \cap V\big) \cup \big(I^+(q) \cap V\big)$. However this set is open, so for large enough n we must have $\xi^{-1}\big(h(y_n), y_n\big) \in \big(I^-(q) \cap V\big) \cup \big(I^+(q) \cap V\big)$, but $\xi^{-1}\big(h(y_n), y_n\big) \in S$. This contradicts S being achronal. $\square$

Corollary 6.5 *An achronal set S is a closed three-dimensional C^0 submanifold of M if and only if* $\text{edge}(S) = \emptyset$. *Hence ∂F (if nonempty) is a closed C^0 submanifold of M for any future set F.*

Proof. By the above proposition, $S \cap \text{edge}(S) = \emptyset$. However, $\text{edge}(S) \subset \overline{S} = S$. Therefore $\text{edge}(S) = \emptyset$. On the other hand suppose $\text{edge}(S) = \emptyset$ and let p be an accumulation point of S. There must exist an open set around U such that every timelike curve around $I^-(p) \cap U$ to $I^+(p) \cap U$ intersects S. If $p \notin S$, then a curve from $I^-(p) \cap U$

to p to $I^+(p) \cap U$ must intersect S (not at p). WLOG assume this point of intersection is the timelike past of p. Let V be a neighborhood of p such that V is in the timelike future of the point of intersection. Since p is an accumulation point V contains a point of S, but this is a contradiction since S is achronal. $\square$

Notice that we can't do much better than C^0 in Theorem 6.4. For example, $J^+(p)$ is not C^1. For a less trivial example, take S to be two disjoint balls in the $t = 0$ hypersurface of Minkowski space. Then $\partial I^+(S)$ will also not be C^1.

Now before continuing, we have to expand our definition of "future directed" from differentiable curves to continuous curves. This is because we will eventually be taking limits of curves, and in general, the limit of such curves will only be continuous. A continuous curve γ is said to be a ***future directed timelike*** (or ***causal***) ***curve*** if for each p in the image of γ, there exists a convex normal neighborhood U of p such that if $\gamma(s_1), \gamma(s_2) \in U$ with $s_1 < s_2$, then there exists a future directed piecewise differentiable timelike (or, respectively, causal) curve in U from $\gamma(s_1)$ to $\gamma(s_2)$.

Let γ be a future directed causal curve. We say $p \in M$ is a ***future endpoint*** if for every neighborhood U of p there exists an s_0 such that $\gamma(s) \in U$ for all $s > s_0$ in the domain γ. ***Past endpoints*** are defined analogously. If γ has no future endpoint (or past endpoint), then γ is said to be ***future inextendible*** (respectively, ***past inextendible***). γ is ***inextendible*** if it's both future and past inextendible. Of course analogous definitions hold for past directed causal curves.

A curve γ is a ***limit curve*** of the sequence $\{\gamma_n\}$ if there is a subsequence $\{\gamma_{n'}\}$ such that for all p in the image of γ, each neighborhood of p intersects all but a finite number

of curves in the subsequence $\{\gamma_{n'}\}$.

We would like to know when a sequence of of curves $\{\gamma_n\}$ will have a limit curve. A sufficient (and necessary) condition is given in Lemma 6.7 below. This lemma is is used multiple times in causal theory, so we give it justice by providing a complete proof. First, the proof of Lemma 6.7 relies on the Arzela-Ascoli theorem which we state as Theorem 6.6:

Theorem 6.6 *Let X be a locally compact Hausdorff space with a countable basis, and let (M, h) be a complete Riemannian manifold with the standard distance function d. Assume that the sequence $\{f_n\}$ of functions from X to M is equicontinuous and that for each $x \in X$, the set $\bigcup_n \{f_n(x)\}$ is bounded with respect to d. Then there exists a continuous function $f : X \to M$ and a subsequence $\{f_{n'}\}$ of $\{f_n\}$ which converges to f uniformly on each compact subset X.*

However, in order to invoke the Arzela-Ascoli theorem, we have to show that continuous causal curves satisfy a certain Lipschitz condition with respect to any auxiliary Riemannian metric h on M.

Let U be a convex normal neighborhood of (M, g) with compact closure $\overline{U}$ contained in a chart V with local coordinates (ct, x, y, z) such that $f = ct : U \to \mathbb{R}$ satisfies the following property: if $q \in I^+(p)$, then $f(p) < f(q)$. By making U small enough, we can find a constant K such that if we define the Lorentzian metric

$$g_1 = -Kd(ct)^2 + dx^2 + dy^2 + dz^2$$

on U, then for all $p \in U$ and $v \in T_pM$, $g(v,v) \leq 0$ implies $g_1(v,v) < 0$. Pictorially, this means the "lightcone of g is smaller than the lightcone of g_1." Let γ be any continuous causal curve (with respect to g) joining $p, q \in U$ with $f(p) < f(q)$. We can parametrize γ by $\gamma(s) = \big(s, x(s), y(s), z(s)\big)$ for all s with $f(p) \leq s \leq f(q)$. Since γ is causal for g, it's causal for g_1, therefore γ satisfies the following Lipschitz condition

$$\sqrt{\sum_a \left[x^a(s_1) - x^a(s_2)\right]^2} \leq \sqrt{1+K}|s_1 - s_2|.$$

where x^a are the components of γ. This Lipschitz condition implies that γ is differentiable almost everywhere and that $\left|\frac{dx^a}{ds}\right| \leq \sqrt{1+K}$ where defined along γ. Thus it makes sense to integrate functions of dx^a/ds along γ.

Now suppose (M, g) is given an auxiliary complete Riemannian metric h with distance function d. Let h_{ab} be the components of h with respect to the coordinates (s, x, y, z). Then the length of γ from s_1 to s_2 with respect to h is

$$L_h(\gamma|_{[s_1,s_2]}) = \int_{s_1}^{s_2} \sqrt{h_{ab}\left(\frac{dx^a}{ds}\right)\left(\frac{dx^b}{ds}\right)}ds.$$

Let $H = \sup\{|h_{ab}(p)| : p \in \overline{U}\}$, then since $|dx^a/ds| \leq \sqrt{1+K}$, we have

$$L_h(\gamma|_{[s_1,s_2]}) \leq 4\sqrt{H}\sqrt{1+K}|s_1 - s_2|.$$

Thus, we can give γ an arc length parametrization with respect to h. Using the paracompactness of M, we can cover (M, g) by a locally finite collection of sets with the properties of U and V above; it follows that we can give *any* causal curve of (M, g) an arc length parametrization with respect to any complete metric h.

Lemma 6.7 *Let $\{\gamma_n\}$ be a sequence of future directed causal curves which are all future inextendible, past inextendible, or inextendible. If p is an accumulation point of $\{\gamma_n\}$,*

then there is a future inextendible (respetively, past inextendible or inextendible) causal curve γ such that p is in the image of γ and γ is a limit curve of $\{\gamma_n\}$.

Proof. We will prove the theorem for inextendible causal curves. The other cases are similar. Let h be an auxiliary complete Riemannian metric for M with distance function d. Give each γ_n an arc length parametrization with respet to h. Then the domain of each γ_n is $\mathbb{R}$ since each curve is inextendible. By shifting parametrizations if necssary, we may find a subsequence $\{\gamma_{n'}\}$ of $\{\gamma_n\}$ such that $\gamma_{n'}(0) \to p$ as $n' \to \infty$ since p is an accumulation point of $\{\gamma_n\}$. Also, since $\{\gamma_{n'}\}$ has an arc length parametrization, we have

$$d\big(\gamma_{n'}(s_1), \gamma_{n'}(s_2)\big) \leq |s_1 - s_2|$$

for each n' and $s_1, s_2 \in \mathbb{R}$. Thus each curve $\gamma_{n'}$ is uniformly continuous so the family $\{\gamma_{n'}\}$ is equicontinuous. Moreover, there exists an integer N such that $d\big(\gamma_{n'}(0), p\big) < 1$ whenever $n' \geq N$. This implies that for each fixed $s \in \mathbb{R}$, the curve $\gamma_{n'}$ restricted to $[-s, s]$ lies in the compact (hence bounded) set $\{q \in M : d(p, q) \leq s + 1\}$ whenever $n' \geq N$. Hence the family $\{\gamma_{n'}\}_{n' \geq N}$ satisfies the hypotheses of the Arezela-Ascoli's theorem, and we thus obtain a continuous curve $\gamma : \mathbb{R} \to M$ and a subsequence $\{\gamma_{n''}\}$ of $\{\gamma_{n'}\}_{n \geq N}$ such that $\{\gamma_{n''}\}$ converges to γ uniformly on each compact subset of $\mathbb{R}$. The convergence $\gamma_{n'}(0)$ implies $\gamma(0) = p$, so all that's left to do is show that γ is causal and inextendible.

We first show γ is causal. Fix $s_1 \in \mathbb{R}$ and let U be a convex normal neighborhood (with respect to the Lorentzian metric g) containing $\gamma(s_1)$. Pick $\delta > 0$ such that the set $V = \{q \in M : d\big(\gamma(s_1), q\big) < \delta\}$ is contained in U. Consider $s_2 \in (s_1, s_1 + \delta)$. For

n'' large enough, the image of the compact set $[s_1, s_2]$ under $\gamma_{n''}$ is completely contained in V. Since $\{\gamma_{n''}(s_1)\} \to \gamma(s_1)$ and $\{\gamma_{n''}(s_2)\} \to \gamma(s_2)$, using the convexity of V, we can find a future directed piecewise differentiable timelike curve from $\gamma(s_1)$ to $\gamma(s_2)$ that lies completely in V. Similarly, if $s_2 \in (s_1 - \delta, s_1)$, then we can find a future directed piecewise differentiable timelike curve from $\gamma(s_2)$ to $\gamma(s_1)$. Thus γ is a future directed causal curve.

Now we show γ is inextendible. Assume otherwise. Then γ has a future endpoint or past endpoint. WLOG assume the former. Then there exists a point $q \in M$ such that $\gamma(s) \to q$ as $s \to \infty$. Let U, V, (ct, x, y, z), and f be as in the discussion above of this lemma such that U is a neighborhood of q. Let $s_0 \in \mathbb{R}$ be such that $\gamma\big([s_0, \infty]\big) \subset U$. The inequality we derived for L_h in the above discussion, implies that a causal curve in U from $f^{-1}\big(f\big(\gamma(s_0)\big)\big)$ to $f^{-1}\big(f(q)\big)$ must have h-length smaller than some number $\delta > 0$. On the other hand, for sufficiently large n'', we must have the image of $[s_0 + 1, s_0 + \delta + 2]$ under $\gamma_{n''}$ is completely contained in $f^{-1}\big(\big[f\big(\gamma(s_0)\big), f(q_0)\big]\big)$. But the h-length of $\gamma_{n''}$ restricted to $[s_0 + 1, s_0 + \delta + 2]$ is $\delta + 1$, which is a contradiction. □

We will use Lemma 6.7 to the prove the following proposition which states that, in general, large portions of achronal boundaries are ruled by null geodesics.

Proposition 6.8 *Let $S \subset M$ be closed. Then each $p \in \partial I^+(S) \setminus S$ lies on a null geodesic contained in $\partial I^+(S)$, which either has a past end point on S, or else is past inextendible.*

Proof. Fix $p \in \partial I^+(S) \setminus S$ and let h be an auxiliary complete Riemannian metric for M. Since $p \in \partial I^+(S)$, there exists a sequence of points $\{p_n\} \subset I^+(S)$ which converge

to p. For each n, let $\gamma_n : [0, s_n] \to M$ be a past directed timelike curve from p_n to $q_n \in S$ and have an arc length parametrization with respect to h. Extend each γ_n to a past inextendible timelike curve $\tilde{\gamma_n} : [0, \infty) \to M$, also with an arc length parametrizaation with respect to h. By lemma 6.7, there exists a subsequence $\{\tilde{\gamma}_{n'}\}$ of $\{\tilde{\gamma}_n\}$ and a continuous past inextendible causal curve γ such that $\gamma(0) = p$ and γ is a limit curve of $\{\tilde{\gamma}_{n'}\}$. By taking another subsequence $\{\tilde{\gamma}_{n''}\}$ of $\{\tilde{\gamma}_{n'}\}$ we can ensure that $s_{n''} \leq s_{m''}$ whenever $n'' < m''$. That is, $\{s_{n''}\}$ is a montonic sequence. We have two cases to consider: (1) $s_{n''} \to s$ for some $s \in (0, \infty)$ or (2) $s_{n''} \to \infty$.

Let's consider case (1). Fix $a \in (0, s)$. Eventually, $s_{n''} > a$, so for n'' large enough, we have $\tilde{\gamma}_{n''}(a) = \gamma_{n''}(a) \in I^+(S)$. So since $\gamma(a) = \lim_{n'' \to \infty} \gamma_{n''}(a)$, it follows that $\gamma(a) \in \overline{I^+(S)}$. Let's suppose $\gamma(a)$ isn't on the boundary, that is $\gamma(a) \in I^+(S)$. Then there exists a point $q \in S$ such that $\gamma(a) \in I^+(q)$ but $p \in J^+\big(\gamma(a)\big)$. Thus we can find a future directed timelike curve from q to p, but this implies $p \in I^+(S)$ which contradicts p being on the boundary. Thus $\gamma(a) \notin I^+(S)$, hence $\gamma(a) \in \partial I^+(S)$. Now since $\partial I^+(S)$ is achronal, no two points of γ can be joined by a timelike curve but they are joined by a causal curve, namely γ. Thus, by approximating γ by piecewise differentiable curves if necessary, proposition 6.1 implies that γ is a null geodesic. Also, since S is closed, $\gamma(s) = \lim_{n'' \to \infty} \gamma_{n''}(s_{n''}) = \lim_{n'' \to \infty} q_{n''} \in S$.

For case (2), the same reasoning as case (1) shows that γ is still a null geodesic. Now if $\lim_{s \to \infty} \gamma(s) \in M$, then this would imply (M, h) is not a complete Rimannian manifold. Thus γ is past inextendible. $\square$

The following lemma will be useful in the next section.

Lemma 6.9 *Let γ be a past inextendible causal curve passing through a point p. Then through any $q \in I^+(p)$, there exists a past inextendible timelike curve λ such that the image of λ is contained in $I^+(\gamma)$.*

Proof. Fix a point r in the image of γ. Using proposition 3.5 and the compactness of the curve γ restricted from p to r, we can find points $r' \in I^+(r)$ and $p' \in I^+(p)$ such that there is a timelike curve λ from r' to p' to q which stays entirely within $I^+(\gamma)$. Now we can continue this process by picking a countable infinite number of points along γ and extending λ appropriately, then λ will be a continuous past inextendible timelike curve from q which is contained in $I^+(\gamma)$. □

6.2 Domains of Dependence and Cauchy Horizons

Given a subset $S \subset M$, $I^+(S)$ physically represents the subset of M which can be influence by physical, massive particles emanating from S. Likewise, $J^+(S)$ physically represents the subset of M which can be influenced by massive particles or light rays (or other forms of radiation along null geodesics) emanating from S. Now we want to consider the points in M which are completely determined by events in a subset $S \subset M$.

Let $S \subset M$ be closed and achronal. We define the ***future domain of dependence of*** S, denoted by $D^+(S)$, as the set of points $p \in M$ such that every past inextendible causal curve through p intersects S.

Intuitively, appropriate knowledge of data on S determines the events in $D^+(S)$, so we can think of S as an "initial condition" for $D^+(S)$. We take S to be closed because,

physically, if we knew the data on a set S, then we would know the data on its closure. Note we have the following inclusions $S \subset D^+(S) \subset J^+(S)$.

$D^-(S)$, the ***past domain of dependence of*** S is defined with "past" replaced by "future." The ***domain of dependence of*** S is $D(S) = D^+(S) \cup D^-(S)$, so $D(S)$ represents the complete set of points in M which are determined by data on S.

The achronality of S shows that points in the timelike past of S can't be in the future domain of dependence of S which is a reasonable property of something we are to consider as an initial condition.

Proposition 6.10 $D^+(S) \cap I^-(S) = \emptyset$.

Proof. Suppose otherwise. Then there exists a $p \in D^+(S) \cap I^-(S)$. So there exists a point $q \in S$ and a future directed causal curve γ from q to p, and there exists a point $r \in S$ and a past directed timelike curve λ from r to p. If γ is timelike, then we contradict achronality because $\gamma \cup \lambda$ is a timelike curve connecting q to r. Otherwise pick a convex normal neighborhood U of p and a point $p' \in I^+(p) \cap U$ such that p' is in the image of λ. Similar to the proof of proposition 6.1, we can use compactness arguments to find a timelike curve γ' which is "close" to γ. Thus we have a future directed timelike curve from q to p' to r which contradicts S being achronal. □

The following proposition shows that the closure of $D^+(S)$ satisfies a similar property used to define $D^+(S)$.

Proposition 6.11 *For a closed and achronal set $S \subset M$, $p \in \overline{D^+(S)}$ if and only if every*

past inextendible timelike curve from p intersects S.

Proof. Define $\widetilde{D}^+(S)$ as the set of points p such that every past inextendible timelike curve from p intersects S. We want to show $\overline{D^+(S)} = \widetilde{D}^+(S)$.

"$\subset$" Clearly $D^+(S) \subset \widetilde{D}^+(S)$, so it suffices to show $\widetilde{D}^+(S)$ is closed. Let $q \in M \setminus \widetilde{D}^+(S)$. Then $q \in M \setminus S$. Since S is closed, there is a neighborhood U of q which does not intersect S. Also, there exists a past inextendible timelike curve γ from q which doesn't intersect S. Let $r \in \gamma \cap U$. Then any point in $I^+(r) \cap U$ contains a past inextendible timelike curve which doesn't intersect S, hence it's an open set around q which is contained in $M \setminus \tilde{D}^+(S)$. Thus $\tilde{D}^+(S)$ is closed.

"$\supset$" Now suppose $p \in \tilde{D}^+(S)$. Either (1) $p \in S$ or (2) $p \in I^+(S)$. In the first case for done. For the first case, we have $S \subset D^+(S) \subset \overline{D^+(S)}$ so we're done. For the second case, let U be any neighborhood about p and let $q \in U \cap I^-(p) \cap I^+(S)$. Suppose we could find a past inextendible causal curve γ from q which didn't intersect S. Then there are two options: (a) γ is contained in $I^+(S)$ or (b) γ intersects $\partial I^+(S)$ at a point $r \notin S$. If (a) is true, then by lemma 6.9 we could find a past inextendible timelike path λ from p whose image is contained in $I^+(\gamma)$. However λ must intersect S, so we violate S being achronal. If (b) is true, then using arguments like those used in lemma 6.9, we can find a timelike path from p to r and extend it arbitrarily into the past. This would then contradict $p \in \tilde{D}^+(S)$. Thus $p \in \overline{D^+(S)}$. □

The interior of domains of dependence have the following property:

Proposition 6.12

(a) $\operatorname{int}\big[D^+(S)\big] = I^-\big[D^+(S)\big] \cap I^+(S)$.

(b) $\operatorname{int}\big[D(S)\big] = I^-\big[D^+(S)\big] \cap I^+\big[D^-(S)\big]$.

Proof. (a) "$\subset$" Suppose $p \in \operatorname{int}\big[D^+(S)\big]$. Let U be convex normal neighborhood around p contained in $D^+(S)$. Let $q \in I^-(p) \cap U$. Since $q \in D + (S)$, there exists a past inextendible causal curve from q which intersects S. Let γ be a past directed timelike curve joining p to q. $\gamma \cup \lambda$ is a causal curve which intersects S which is not a null geodesic, so by proposition 6.1, $p \in I^+(S)$. Let $r \in I^+(p) \cap U$. Since $r \in D^+(S)$, we have $p \in I^-\big[D^+(S)\big]$.

"$\supset$" Suppose $p \in I^-\big[D^+(S)\big] \cap I^+(S)$. Since this intersection is an open set, there is a neighborhood U of p contained in the intersection. Fix $q \in U$ and suppose there exists a past inextendible causal curve γ from q which does not intersect S. Since $q \in I^-\big[D^+(S)\big]$, there exists a point $r \in I^+(q)$ such that $r \in D^+(S)$. Let γ be a past directed timelike path from r to q. If γ intersected S, then $q \in I^+(S)$ contradicts S being achronal. Therefore $\gamma \cup \lambda$ is a past inextendible causal curve which does not intersect S. This contradicts $r \in D^+(S)$. Thus $U \subset D^+(S)$.

(b) "$\subset$" Let $p \in \operatorname{int}\big[D(S)\big]$. There exists a neighborhood U of p contained in $D^+(S) \cup D^-(S)$. WLOG suppose $p \in D^+(S)$. Let $q \in I^+(p) \cap U$. q is either in $D^+(S)$ or $D^-(S)$. If q is in the latter, then we can find a timelike curve from S to p to q to S which contradicts S being achronal. Therefore $q \in D^+(S)$ which implies $p \in I^-\big[D^+(S)\big]$. Since $p \in D^+(S)$, there is a past directed timelike curve from p to S. Since $S \subset D^-(S)$, we have $p \in I^+\big[D^-(S)\big]$, also. Thus $p \in I^-\big[D^+(S)\big] \cap I^+\big[D^-(S)\big]$.

"$\supset$" Let $p \in I^-\big[D^+(S)\big] \cap I^+\big[D^-(S)\big]$. Since this set is open, there is a neighborhood

U of p contained in the intersection. Let $q \in U$. Suppose, just maybe, $q \notin D(S)$. Then there exists a past inextendible causal curve γ_1 and a future inextendible causal curve γ_2, both starting at q, such that neither γ_1 nor γ_2 intersect S. Now since $q \in I^-\big[D^+(S)\big]$, there exists a point $r_1 \in D^+(S)$ and a past directed timelike curve λ_1 from r_1 to q. Likewise, we can find a point $r_2 \in D^-(S)$ and a future directed timelike curve λ_2 from r_2 to q. Now if λ_1 didn't intersect S, then $\lambda_1 \cup \gamma_1$ is a past inextendible timelike curve from r_1 which doesn't intersect S; this contradicts $r_1 \in D^+(S)$. Therefore λ_1 must intersect S. Likewise λ_2 must intersect S. But S is achronal, so we must have λ_1 and λ_2 intersect S at q which implies $q \in S \subset D(S)$, contradicting our initial assumption $q \notin D(S)$. Therefore $U \subset D(S)$. □

For a closed and achronal set S, we define the ***future Cauchy horizon of*** S, denoted $H^+(S)$, by

$$H^+(S) = \overline{D^+(S)} \setminus I^-\big[D^+(S)\big].$$

We define the ***past Cauchy horizon of*** S, $H^-(S)$, analogously. The ***Cauchy horizon of*** S is simply $H(S) = H^+(S) \cup H^-(S)$. Physically, $H^+(S)$ marks the limit of M controlled by S. The adjective "Cauchy" will be come clear in section 6.4.

$H^+(S)$ is closed since it's the intersection of two closed sets, $\overline{D^+(S)}$ and $\big(M \setminus I^-\big[D^+(S)\big]\big)$. Also $H^+(S)$ is achronal. To see this, notice that

$$I^-\big[H^+(S)\big] \subset I^-\big[\overline{D^+(S)}\big] = I^-\big[D^+(S)\big] \subset M \setminus H^+(S).$$

Thus $I^-\big[H^+(S)\big] \cap H^+(S) = \emptyset$ which implies $H^+(S)$ is achronal. $H^+(S)$ looks like a boundary and indeed it is:

Proposition 6.13 $H^+(S) = \partial I^-\big[D^+(S)\big] \cap \big[I^+(S) \cup S\big]$.

Proof. "$\subset$ " Suppose $p \in H^+(S)$. Let U be any neighborhood about p. Since $p \in \overline{D^+(S)}$, there is a point $q \in U$ such that $q \in D^+(S)$. Let γ be a past directed timelike curve from q to S. Let $r \neq q$ be a point on γ which is also contained in U. Then $r \in I^-\big[D^+(S)\big]$. Therefore $p \in \overline{I^-\big[D^+(S)\big]}$ but we also have $p \notin I^-\big[D^+(S)\big]$, so $p \in \partial I^-\big[D^+(S)\big]$. Now assume $p \notin I^+(S)$. Then every past directed timelike curve from p doesn't intersect S. Let $\{U_n\}$ be a shrinking sequence of convex normal neighborhoods about p and let $q_n \in U_n \cap D^+(S)$. Every past inextendible causal curve from q_n must intersect S, but for n large enough, the past inextendible causal curves from q_n will start intersecting the past inextendible timelike curves from p. Therefore we must have $q_n \in S$ for n larger than some integer N. But S is closed, so $p = \lim_{n\to\infty} q_n \in S$.

"$\supset$" Suppose $p \in \partial I^-\big[D^+(S)\big] \cap \big[I^+(S) \cup S\big]$. By definition of boundary, $p \notin I^-\big[D^+(S)\big]$. Let U be any neighborhood of p. First assume $p \in I^+(S)$. There exists a point $q \in U \cap I^+(S)$ such that $q \in I^-\big[D^+(S)\big]$. The combination $q \in I^-\big[D^+(S)\big] \cap I^+(S)$ implies $q \in D^+(S)$. Therefore $p \in \overline{D^+(S)}$. If $p \notin I^+(S)$, then $p \in S \subset D^+(S) \subset \overline{D^+(S)}$. $\square$

Another useful fact about future Cauchy horizons is that they share their edges with S.

Proposition 6.14 *For S closed and achronal, we have $I^+\big[\mathrm{edge}(S)\big] \cap \overline{D^+(S)} = \emptyset$ and* $\mathrm{edge}\big[H^+(S)\big] = \mathrm{edge}(S)$.

Proof. For the first part let $p \in \text{edge}(S)$. Let $q \in I^+(p)$. There is a neighborhood U of s in $I^+(p)$. For any neighborhood V about p, there are points $r \in I^-(p) \cap U$ and $s \in I^+(p) \cap U$ and a future directed timelike curve γ from s to r which doesn't intersect S. Now choose V small enough so that $s \in I^-(U)$. Thus we can find a timelike curve for any point in U to s which continues along γ to r and extends to the past indefinitely. This curve will not meet S because S is achronal, so by proposition 6.11 $q \notin \overline{D^+(S)}$.

For the second part take $p \in \text{edge}(S)$. Since S is closed, we have $p \in S \subset \overline{D^+(S)}$. The first part of this proposition implies $p \notin I^-\big[\overline{D^+(S)}\big] = I^-\big[D^+(S)\big]$. Therefore $p \in H^+(S)$. To see that in fact $p \in \text{edge}\big[H + (S)\big]$, let U be a neighborhood of p. There are points $q \in I^-(p) \cap U$ and $r \in I^+(p) \cap U$ and a future directed timelike curve γ from q to r not meeting S. γ also can't meet $H^+(S)$ because every past directed inextendible timelike curve from $H^+(S)$ intersects S which is a contradiction since S is achronal and $q \in I^-(S)$. I am having difficulty proving the converse so it's left as an exercise.

Similar to proposition 6.8, a large portion of a future Cauchy horizon is ruled by null geodeiscs.

Theorem 6.15 *Every point $p \in H^+(S)$ lies on a null geodesic contained entirely within $H^+(S)$ which either is past inextendible or has a past endpoint on the edge of S.*

Proof. We can consider the case $p \notin \text{edge}(S)$ because otherwise the trivial curve $\gamma(s) = p$ for all s is a null geodesic with a past endpoint on the edge of S. Therefore either (1) $p \in I^+(S)$ or (2) $p \in S \setminus \text{edge}(S)$.

Assume (1). Since $p \notin I^-\big[D^+(S)\big]$, for every $q \in I^+(p)$, there exists a past inex-

tendible causal curve from q which does not intersect S. Let $\{q_n\}$ be a sequence of points in $I^+(p)$ which converges to p and $\{\gamma_n\}$ a sequence of past inextendible causal curves starting at q_n which do not intersect S. Since p is an accumulation point of $\{\gamma_n\}$, by lemma 6.7 there is a past inextendible causal curve γ such that γ is a limit curve of $\{\gamma_n\}$ and p is in the image of γ.

We will show $\gamma \cap I^+(S)$ is a null geodesic. Suppose γ entered $\operatorname{int}\big[D^+(S)\big]$. Then γ_n would enter $D^+(S)$ for sufficiently large n which contradicts our construction of $\{\gamma_n\}$. By proposition 6.12 (a), γ does not enter $I^+(S) \cap I^-\big[D^+(S)\big]$. Since $I^-(p) \subset I^-\big[\overline{D^+(S)}\big] = I^-\big[D^+(S)\big]$, $\gamma \cap I^+(S)$ is a past directed causal curve from p which does not enter $I^-(p)$. Therefore $\gamma \cap I^+(S)$ is a null geodesic by proposition 6.1.

Now we show $\gamma \cap I^+(S) \subset H^+(S)$. We already know $\gamma \nsubseteq \operatorname{int}\big[D^+(S)\big] = I^+(S) \cap I^-\big[D^+(S)\big]$. Therefore $\big[\gamma \cap I^+(S)\big] \cap I^-\big[D^+(S)\big] = \emptyset$, so it suffices to show $\gamma \cap I^+(S) \subset \overline{D^+(S)}$. Using proposition 6.11, suppose a past inextendible timelike curve from some point of $\gamma \cap I^+(S)$ failed to intersect S, then using compactness arguments and proposition 3.5, we can find a past inextendible timelike curve from p which does not intersect S, but this contradicts $p \in \overline{D^+(S)}$ by proposition 6.11. Therefore $\gamma \cap I^+(S) \subset H^+(S)$.

Now since $p \in I^+(S)$, we have found a *nontrivial* null geodesic $\lambda \subset \gamma \cap I^+(S)$ from p which remains in $H^+(S)$. Now extend λ to an inextendible null geodesic $\tilde{\lambda}$. If $\tilde{\lambda}$ leaves $H^+(S)$, then let $\tilde{\lambda}_1$ be the portion of $\tilde{\lambda}$ which remains in $H^+(S)$ and let r be the past endpoint of $\tilde{\lambda}_1$. But this implies $r \in \operatorname{edge}\big(H^+(S)\big) = \operatorname{edge}(S)$.

Now assume (2). Since $p \notin \operatorname{edge}(S)$, we can find an open set U around p such that no causal curve contained in U from some point in $I^+(p) \cap U$ can enter $I^-(p) \cap U$ without intersecting S. Therefore the same argument used for (1) shows that we can find

a *nontrivial* past directed null geodesic from p which either remains in $H^+(S)$ or has an endpoint on edge(S). $\square$

The Cauchy horizon of a closed an achronal set S is precisely the boundary of the domain of dependence of S.

Proposition 6.16 $H(S) = \partial D(S)$.

Proof. " $\subset$ " We have

$$\begin{aligned} H(S) &= \big(\overline{D^+(S)} \setminus I^-\big[D^+(S)\big]\big) \cup \big(\overline{D^-(S)} \setminus I^+\big[D^-(S)\big]\big) \\ &\subset \big(\overline{D^+(S)} \cup \overline{D^-(S)}\big) \setminus \big(I^-\big[D^+(S)\big] \cap I^+\big[D^-(S)\big]\big) \\ &= \overline{D(S)} \setminus \operatorname{int}\big[D(S)\big] \\ &= \partial D(S). \end{aligned}$$

"$\supset$" On the other hand, assume $p \in \overline{D(S)} \setminus \operatorname{int}\big[D(S)\big]$. $p \in \overline{D^+(S)} \cup \overline{D^-(S)}$. Suppose $p \in \overline{D^+(S)}$. We want to show $p \notin I^-\big[D^+(S)\big]$. If this is not the case, then since $p \notin \operatorname{int}\big[D(S)\big] = I^-\big[D^+(S)\big] \cap I^+\big[D^-(S)\big]$, we must have $p \notin I^+\big[D^-(S)\big]$ so $p \notin I^+(S)$, but this contradicts $p \in \overline{D^+(S)}$ by proposition 6.11. Therefore $p \in H^+(S)$. Likewise, if we assumed $p \in \overline{D^-(S)}$, then we would have found $p \in H^-(S)$. $\square$

6.3 Causality Conditions

In this section, we discuss reasonable causality conditions that might hold in our universe. For example, observers on closed timelike curves could alter their past leading to examples

of the "grandfather paradox." The spacetime (M, g) satisfies the ***chronology condition at*** p if M contains no closed timelike curves through p. (M, g) satisfies the ***chronology condition*** if it satisfies the chronology condition at every $p \in M$. The following theorem essentially eliminates the consideration of compact spacetimes as physical models for our universe because they don't satisfy the chronology condition.

Theorem 6.17 *If (M, g) is compact, then (M, g) contains a closed timelike curve.*

Proof. The set $\{I^+(p) : p \in M\}$ is an open cover for M so we can extract a finite subcover $\{I^+(p_1), \ldots, I^+(p_k)\}$. We can assume this is the minimal number of such sets covering M. We must have $p_1 \in I^+(p_i)$ for some i. But since $I^+(p_1) \subset I^+(p_i)$, we must have $i = 1$ otherwise we wouldn't have a minimal subcover. Therefore $p_1 \in I^+(p_1)$ so there exists a closed timelike curve through p_1. □

A slightly stronger condition is the ***causality condition*** which states that (M, g) contains no closed causal curves. There exist spacetimes which satisfy the causality condition but have curves which come arbitrarily close to violating the causality condition. Thus a small perturbation of the metric g could cause (M, g) to violate the causality condition. To avoid this type of scenario, we will consider strongly causal spacetimes. (M, g) is said to be ***strongly causal at*** $p \in M$ if every neighborhood U of p, there exists a neighborhood V of p contained in U such that no causal curve intersects V more than once (i.e. no causal curve intersects V on disconnected sets). (M, g) is ***strongly causal*** if it's strongly casual at every $p \in M$.

Compact sets within strongly causal spacetimes are known to "imprison" causal

curves confined within them.

Proposition 6.18 *Suppose (M, g) is strongly causal and let $K \subset M$ be compact. Then every causal curve γ whose image is contained in K must have past and future endpoints in K.*

Proof. If the domain of γ is a closed interval, then since γ is confined within C, it must have past and future endpoints in K. Thus we can assume the domain of γ is $\mathbb{R}$. We will show γ has a future endpoint in K, the past endpoint can be shown analogously. Let $\{s_n\}$ be an increasing sequence of numbers which diverges to infinity and set $p_n = \gamma(s_n)$. Then $\{p_n\}$ is a sequence in K and thus it has an accumulation point $p \in K$. Suppose p is not the future endpoint of γ. Then there exists a neighborhood U of p such that for any $s_1 \in \mathbb{R}$ there exists an $s_2 > s_1$ such that $\gamma(s_2) \notin U$. This condition must also hold true for any open set $V \subset U$, but this means γ intersects V more than once, since infinitely many points of the sequence $\{\gamma(s_n)\}$ enter V but γ never remains in V. This implies (M, g) is not strongly causal which is a contradiction. Hence $p \in K$ is a future endpoint of γ. □

Our goal now is to find necessary and sufficient conditions for a spacetime to be strongly causal at a point. The following terminology will be helpful. Let U be an open set in M, we define $\langle p, q\rangle_U$ as the set of points $r \in M$ such that there is a future directed timelike curve from p to r to q lying completely within U. We put $\langle p, q\rangle = \langle p, q\rangle_M$. Notice that $\langle p, q\rangle = I^+(p) \cap I^-(q)$.

Proposition 6.19 *Let U be a convex normal neighborhood and pick $p, q \in U$, then any causal curve lying in U cannot intersect $\langle p, q\rangle_U$ more than once (i.e. it doesn't intersect $\langle p, q\rangle_U$ on disconnected sets).*

Proof. Let γ be any causal curve lying in N. Let $r, s \in \gamma \cap \langle p, q\rangle_U$ with $s \in J^+(r)$. Since U is convex there are future directed timelike geodesics from p to r and from s to q. Thus for any point t between r and S along γ, we can find a timelike curve (using proposition 6.1 if necessary) from x to t to y which, by definition, must remain in $\langle x, y\rangle_U$. Since this is true for all $r, s \in \gamma \cap \langle p, q\rangle_U$, $\gamma \cap \langle p, q\rangle_U$ can't have more than one connected component. □

The following lemma is a certain converse to proposition 6.1.

Proposition 6.20 *Let U be a convex normal neighborhood and let V be an open set in U, then for any $p \in V$ there exist $q, r \in V$ such that $p \in \langle q, r\rangle_U \subset V$.*

Proof. Let (ct, x, y, z) be normal coordinates for N with origin at p. Choose $\epsilon > 0$ so that the normal coordinate ball, given by $(ct)^2 + x^2 + y^2 + z^2 < \epsilon^2$, is contained in V and such that $\partial/\partial t$ is future directed. Take q to be the point at $(-\frac{1}{2}\epsilon, 0, 0, 0)$ and r to be the point at $(\frac{1}{2}\epsilon, 0, 0, 0)$. Let $s \in \langle q, r\rangle_U$. There exists a timelike curve γ from q to s and extend γ within N. γ meets the hemisphere defined by $(ct)^2 + x^2 + y^2 + z^2 < \epsilon^2$, so γ must intersect the light cone of past null geodesics emanating from r. Let us call u this point of intersection and λ the null geodesics joining u to r. Notice that there can be no point v to the future of u on γ which is in the past of r, because otherwise we could find

a timelike curve from u to r but this contradicts λ is a null geodesic from u to r. Thus $w \in B \subset V$. □

Corollary 6.21 *If U is a convex normal neighborhood of (M, g), then (U, g) is a strongly causal spacetime.*

Proof. This follows immediately from proposition 6.19 and 6.20. □

Here what come at our first necessary and sufficient for a spacetime to be strongly causal at a point. U is said to be a ***local causality neighborhood*** if no causal curve intersects U more than once and $\overline{U}$ is compact contained in a convex normal neighborhood.

Theorem 6.22 *M is strongly causal at p if and only if p is contained in some local causality neighborhood.*

Proof. "⇒" Since M is strongly casual at p, we can find arbitrarily small local causality neighborhoods around p.

"⇐" Suppose p belongs to a local causality neighborhood U whose closure is contained in a convex normal neighborhood V. By proposition 6.20, we can find arbitrarly small neighborhoods about p which are of the form $p \in \langle q, r\rangle_V \subset U$. If a causal curve γ were to intersect $\langle q, r\rangle_V$ more than once, then by proposition 6.19, γ must leave and reenter V which implies it leaves and reenters U which cannot be since U is a local causality neighborhood. □

Corollary 6.23 *The set of points at which M is strongly causal is an open set.*

The next proposition is the motivation for why we choose local causality neighborhoods to have compact closure within convex normal neighborhoods.

Proposition 6.24 *If U is a local causality neighborhood, then any future (or past) inextendible causal curve γ cannot be completely contained in U.*

Proof. Suppose, just maybe, γ is a future inextendible causal curve in U. γ can't be a closed because otherwise M would not be strongly causal at p. So let $\{p_n\}$ be a sequence of distinct points proceeding points along γ with the property that if $q \in \gamma$, then there exists some n such that p_n lies to the future of q on γ. Since $\overline{U} \subset V$ for some convex normal neighborhood V, there exists an accumulation point $p \in \overline{U}$ of $\{p_n\}$. Since p is not a future endpoint of γ, there exists a neighborhood O of p such that there are points arbitrarily far into the future along γ not contained in O. Using proposition 6.20, pick $q, r \in O$ so that $p \in \langle q, r \rangle_V \subset O$. $\langle q, r \rangle_V$ contains infinitely many points of $\{p_n\}$ on γ but γ also fails to contain infinitely many points on γ between the p_n's. This implies γ must leave and renter $\langle q, r \rangle_V$ which contradicts proposition 6.19. □

If a point $p \in M$ doesn't satisfy the causality condition, then it certainly doesn't satisfy the strong causality condition. But M satisfies the causality condition? What kind kind of behavior can we find at p that can explain strong causality violation without violating causality somewhere? Theorem 6.26 will answer this question for us but first we need a lemma.

Lemma 6.25 *Strong causality fails at $p \in M$ if and only if there exists a point $q \in J^-(p)$ with $q \neq p$ such that the following condition holds: if $p \in I^+(r)$ and $s \in I^+(q)$, then $s \in I^+(r)$.*

Proof. "$\Rightarrow$" Suppose strong causality fails at p. Let V be a convex normal neighborhood containing p with compact closure and let $U_n = \langle q_n, r_n \rangle_V$ with $\overline{U_n} \subset V$ be a nested sequence of neighborhoods of p converging to p (i.e. $U_{n+1} \subset U_n$ and $\{p\} = \bigcap_n U_n$). Each U_n can't be a causally convex neighborhood because that would contradict theorem 6.22. Therefore there exists a causal curve γ_n which intersects U_n more than once. By proposition 6.19, each γ_n cannot completely lie in V. We can take γ_n to have a past endpoint $a_n \in U_n$ and to exit V first at $b_n \in \partial V$, finally to reenter V at c_n and to terminate with future endpoint d_n. Since ∂V is a closed subset of the compact set $\overline{V}$, there is a point $c \in \partial V$ which is an accumulation point of $\{c_n\}$. We can find a future directed causal curve from c_n to d_n. By extending these curves, we can use lemma 6.7 to find a future directed causal curve λ from c to p. Choose $q \in \lambda$. Now suppose $p \in I^+(r)$ and $s \in I^+(q)$. Since $p \in I^+(r)$ which is open, there is an N_1 such that $n > N_1$ implies $U_n \subset I^+(r)$. Hence $a_n \in I^+(r)$ whenever $n > N_1$. Likewise, $c \in I^-(s)$ so there exists an N_2 such that $c_n \in I^-(y)$ whenever $n > N_2$. Thus, whenever $n > \max\{N_1, N_2\}$ we can find a timelike curve from r to a_n to b_n to c_n to s which implies $s \in I^+(r)$.

"$\Leftarrow$" Assume $q \in J^-(p)$ with $q \neq p$ and that $p \in I^+(r)$ and $s \in I^+(q)$ imply $s \in I^+(r)$. Let U and V be disjoint neighborhoods around p and q, respectively. Fix $a \in U \cap I^+(p)$ and let $r \in U \cap I^-(p)$ and $s \in V \cap I^+(q) \cap I^-(a)$ (this set is nonempty because $q \in I^-(a)$). Thus, we can find a causal curve from r to s to a which must

intersect U more than once. Therefore U cannot be a local causal neighborhood. Since U was arbitrary, Theorem 6.22 implies M is not strongly causal at p. $\square$

Theorem 6.26 *If strong causality fails at $p \in M$ but M satisfies the causality condition, then there is an inextendible null geodesic γ through p such that every point on γ violates strong causality and if u and v are any two points of γ with $v \in J^+(u)$ and $u \neq v$, then $v \in I^+(r)$ and $s \in I^+(u)$ together imply $s \in I^+(r)$.*

Proof. Like in the proof of lemma 6.25, let V be a convex normal neighborhood containing p with compact closure and let $U_n = \langle q_n, r_n \rangle_V$ with $\overline{U_n} \subset V$ be a nested sequence of neighborhoods of p converging to p. Let γ_n, a_n, b_n, c_n, and d_n be as in the proof of lemma 6.25. Let (b, c) be an accumulation point of $\{(b_n, c_n)\} \subset \partial V \times \partial V$ (which is compact). Let λ_1 and λ_2 be the future directed causal geodesics from p to b and c to p, respectively. There are a variety of situations that can occur:

If either λ_1 or λ_2 are both timelike, then for n large enough we can find $b_n \in I^+(p)$ and $c_n \in I^-(p)$ so there would be a closed timelike curve through p which contradicts the causality of M.

If λ_1 is timelike and λ_2 is null, then we could find a $q \in \langle p, b \rangle_N$. $x \in I^+(c)$ so for large enough n we can find points c_n and b_n such that $x \in I^+(c_n)$ and $b_n \in I^+(x_n)$. Thus there is a closed timelike curve through x which contradicts the causality of M. Likewise, we cannot have λ_1 null and λ_2 timelike.

If λ_1 and λ_2 are null but $\lambda_1 \cup \lambda_2$ is not C^1 at p, then $b \in I^+(c)$ by proposition 6.1. So again, there is some large n such that for any $q \in \langle c, b \rangle_V \neq \emptyset$, we can find a timelike

curve from q to b_n to c_n to q, again contradicting the causality of M.

Therefore we must have $\lambda_1 \cup \lambda_2$ is a null geodesic. We will first show that strong causality is violated everywhere along of $\lambda_1 \cup \lambda_2$. Let $w \neq c$. Fix r and s such that $w \in I^+(r)$ and $s \in I^+(c)$. By lemma 6.25, it suffices to show $s \in I^+(r)$. Indeed for large enough n, we can find a timelike curve from r to b_n to c_n to s. For $w = c$, consider the time dual statement of lemma 6.25. Since we know b is strongly causal, time duality implies c is also strongly causal.

Now let λ be the inextendible null geodesic which contains $\lambda_1 \cup \lambda_2$. Since b is strongly causal we repeat the whole process for b (in place of p) so there is a null geodesic $\lambda_1' \cup \lambda_2'$ from some point c' to b to b'. Like before, $\lambda_1' \cup \lambda_2'$ must be a null geodesic but $\lambda_1 \subset \lambda_1' \cup \lambda_2'$, therefore $\lambda_1' \cup \lambda_2' \subset \lambda$. By continuing this construction into the future and into the past, we see that λ is an inextendible null geodesic through p such that every point of λ violates strong causality. Now let u and v be any two points of λ with $v \in J^-(u)$ and $u \neq v$. Since the portion of γ from u to v is compact, we can cover it by a finite number of convex normal neighborhoods with compact closure $\{V_1, \ldots, V_k\}$. Using the points which λ intersects ∂V_i, we can find a timelike curve from r to s for any r and s which satisfy $v \in I^+(r)$ and $s \in I^+(u)$. □

6.4 Cauchy Surfaces and Globally Hyperbolic Spacetimes

If S is a closed and achronal set such that $D(S) = M$, then S is called a ***Cauchy surface***. Equivalently, S is a Cauchy surface if every inextendible causal curve intersects

S, so a cauchy surface can be thought of as an instant in time which determines the conditions of the universe. This is why these sets are called "Cauchy;" they describe the initial conditions of the universe. It is clear that a Cauchy surface S can't have an edge since S is achronal. Therefore corollary 6.5 implies S is a three-dimensional C^0 submanifold of M. This justifies the term "surface." Notice that a closed, achronal set $S \subset M$ is always a Cauchy surface for the spacetime $\big(\mathrm{int}\big[D(S)\big], g\big)$. The $t = \text{const}$ hypersurfaces in Minkowski space and $\tau = \text{const}$ hypersurfaces in the FRW solutions are Cauchy surfaces. Also, the maximal extension of the Schwarzschild solution admits a Cauchy surface.

If S is not a Cauchy surface for (M, g), then the points in the Cauchy horizon $H^+(S)$, which is intuitively the set of points which can't be determined by S, is called a "horizon." In fact, we have the following

Proposition 6.27 *A nonempty closed and achronal set S is a Cauchy surface for (M, g) if and only if $H(S) = \emptyset$.*

Proof. "$\Rightarrow$" Suppose $D(S) = M$. Recall that we take our spacetimes to be connected so $D(S)$ is both open and closed. From proposition 6.16, we have $H(S) = \partial D(S) = \overline{D(S)} \setminus \mathrm{int}\big[D(S)\big] = \emptyset$.

"$\Leftarrow$" Suppose $H(S) = \emptyset$. Then $\overline{D(S)} = \mathrm{int}\big[D(S)\big]$. Therefore $D(S)$ is both open and closed so it equals M. $\square$

A useful criterion for Cauchy surfaces is given by proposition 6.29 below for which we need the following obvious, but slightly difficult to prove, lemma.

Lemma 6.28 *If S is a Cauchy surface and γ is an inextendible causal curve, then γ intersects S, $I^+(S)$, and $I^-(S)$.*

Proof. That γ intersects S is in the definition of Cauchy surface. Suppose γ did not intersect $I^-(S)$, then $\gamma \subset S \cup I^+(S)$ since $D(S) = M$ implies $S \cup I^+(S) \cup I^-(S) = M$. By lemma 6.9, we can find a past inextendible timelike curve λ such that

$$\lambda \subset I^+(\gamma) \subset I^+\big(S \cup I^+(S)\big) = I^+(S).$$

If λ intersected S at p, then pick $q \in \lambda \cap I^-(p)$. Since $q \in I^+(S)$, we can find a point $r \in S$ and a future directed time like curve from r to q to p which is a contradiction since S is achronal. Therefore $\lambda \cap S = \emptyset$. So if we extend λ into a future inextendible timelike curve, it must intersect S at some point p. But again, since $\lambda \subset I^+(S)$ and S is achronal, this cannot be. Therefore γ must intersect $I^-(S)$. In a similar way, γ intersects $I^+(S)$. □

Proposition 6.29 *If S is closed, achronal, and without edge, then S is a Cauchy surface if and only if every inextendible null geodesic intersect intersects S, $I^+(S)$, and $I^-(S)$.*

Proof. "⇒" This is just a special case of lemma 6.28.

"⇐" Suppose S is not a Cauchy surface. By proposition 6.27 either $H^+(S) \neq \emptyset$ or $H^-(S) \neq \emptyset$. WLOG let's say $H^+(S) \neq \emptyset$. Since $\text{edge}(S) = \emptyset$, by theorem 6.15 there exists a past inextendible null geodesic γ which remains forever in $H^+(S)$. Therefore $\gamma \nsubseteq I^-\big[D^+(S)\big]$ which implies $\gamma \nsubseteq I^-(S)$. If we extend γ into the future, making it into an inextendible null geodesic, it still cannot enter $I^-(S)$ since S is achronal. □

If the spacetime (M, g) had a Cauchy surface S, then it's easy to see that the chronology condition holds on (M, g). Indeed if γ was a closed timelike curve, we could make it inextendible by going around and around itself. Then γ would have to intersect S but this would violate the achronality of S. In fact using the above proposition we can show that (M, g) satisfies the causality condition. But we can do even better:

Proposition 6.30 *Let S be a Cauchy surface for (M, g). Then (M, g) is strongly causal.*

Proof. Suppose strong causality were violated at $p \in M$. Since (M, g) satisfies the causality condition, by theorem 6.26 there is an inextendible null geodesic through p with the property that if $v \in \gamma \cap J^+(u)$ with $u \neq v$, then for every r and s which satisfy $v \in I^+(r)$ and $s \in I^+(u)$, we have $s \in I^+(r)$. By lemma 6.28, γ must intersect $I^-(S)$, let's say at u, and γ must intersect $I^+(S)$, let's say at v. Then we can find a future directed timelike curve from $s \in I^+(u) \cap I^-(S)$ to $r \in I^-(v) \cap I^+(S)$ which contradicts the achronality of S. □

Lastly, any two Cauchy surfaces are homeomorphic.

Proposition 6.31 *Let S and S' be two Cauchy surfaces for (M, g), then S and S' are homeomorphic.*

Proof. Since (M, g) is time orientable, there is a nonvanishing vector field v on M. Each integral curve of v must have precisely one intersection point with S and S'. Thus we can find a one to one map from S onto S' obtained by identifying points the

intersection points of the integral curves of v. The continuity of v implies this map is continuous, and the inverse is continuous since it's just the time dual. $\square$

Now we define globally hyperbolic spacetimes. Their relation to Cauchy surfaces will be given in the following theorems below. We will find that globally hyperbolic spacetimes are equivalent to those spacetimes admitting a Cauchy surface.

A spacetime (M, g) is said to be ***globally hyperbolic*** provided that (1) M is strongly causal and (2) $J^+(p) \cap J^-(q)$ are compact for all $p, q \in M$. The reason for the name "globally hyperbolic" is that the wave equation for a δ function at $p \in M$ has a unique solution in a globally hyperbolic spacetime. Intuitively, the second condition says that there are no holes or gaps in M. Recall that we constructed spacetimes with holes in them to show $J^+(p)$ isn't necessarily closed. This can't be in globally hyperbolic spacetimes.

Proposition 6.32 *Let (M, g) be globally hyperbolic and fix a compact set A. Then $J^+(A)$ and $J^-(A)$ are closed in M.*

Proof. Fix $p \in \overline{J^+(A)}$. Let $\{p_n\}$ a sequence of points in $J^+(A)$ converging to p. There exist points $q_n \in A$ and future directed causal curves γ_n connecting q_n to p_n. Let $r \in I^+(p)$. Since A is compact, there is a point $q \in A$ which is an accumulation point of $\{q_n\}$. By lemma 6.7 there is a future directed limit curve γ of $\{\gamma_n\}$ which passes through q. Let U be any neighborhood of p, then U contains some all p_n for n large enough. Then using γ, we can find a causal curve from q to p_n for n sufficiently large. Therefore

we have $p \in \overline{J^+(q)}$. Now let $r \in I^+(p)$ so that $p \in \overline{J^-(r)}$. Thus we have

$$p \in \overline{J^+(q)} \cap \overline{J^-(r)} = \overline{J^+(q) \cap J^-(r)} = J^+(q) \cap J^-(r).$$

The last equality follows because since $J^+(q) \cap J^-(r)$ is compact, it's also closed. Thus $p \in J^+(q) \subset J^+(A)$ which shows $J^+(A)$ is closed. Likewise $J^-(A)$ is closed. □

Theorem 6.33 *If S is a Cauchy surface for (M, g), then (M, g) is globally hyperbolic.*

Proof. By proposition 6.30, we know (M, g) is strongly causal. Fix $p, q \in M$. We want to show $J^+(p) \cap J^-(q)$ is compact, so let $\{p_n\}$ be an infinite sequence of points in $J^+(p) \cap J^-(q)$. We want to find an accumulation point $\{p_n\}$. For each n, let γ_n be a future directed causal curve from p to p_n to q. We will show that there is a future directed causal limit curve of $\{\gamma_n\}$ which goes from p to q. Notice that each γ_n are future inextendible in the spacetime $(M - \{q\}, g)$, so lemma 6.7 implies there is a future directed causal limit curve γ passing through p and is inextendible in $(M - \{q\}, g)$. WLOG assume $p \in D^-(S)$. There are two possibilities: (1) $q \in D^-(S)$ or (2) $q \in I^+(S)$. If (1) is true, then γ cannot enter $I^+(S)$ because $q \notin I^+(S)$. Therefore by lemma 6.28, γ can't be future inextendible in (M, g), so it has a future endpoint in (M, g). Either q is the future endpoint of γ or γ extends beyond q. In the latter case, using the fact that γ is a limit curve of $\{\gamma_n\} \subset J^+(p) \cap J^-(q)$, we could find a closed causal curve. Therefore we must have q is the future endpoint of γ. If (2) were true, then it is sufficient to consider $p \in I^-(S)$, as otherwise $p \in S \subset D^+(S)$ and so the time dual of (1) applies. In this case γ does enter $I^+(S)$ since $q \in I^+(S)$. Fix $r \in \gamma \cap I^+(S)$. Now the points q and r fall

in the time dual possibility of (1) so we can find a past directed causal curve from q to r and then to p using γ. In conclusion, we have found a future directed causal curve γ from p to q which is a limit curve of $\{\gamma_n\}$.

Using Theorem 6.22, cover γ by local causality neighborhoods and choose a finite subcover $\{U_1, \ldots, U_k\}$. Let $U = \bigcup_{i=1}^{k} U_i$. Then $\overline{U}$ is compact and contains γ. We must have $\gamma_n \subset U$ for infinitely many n otherwise we contradict γ being a limit curve. Therefore there are an infinite number of $\{p_n\}$ in U, so there is an accumulation point $\tilde{p} \in \overline{U}$ of $\{p_n\}$. By using local causality neighborhoods, we have ensured that $p \in \gamma \subset J^+(p) \cap J^-(q)$. $\square$

The converse of Theorem 6.33 is true also, but in order to show it we need to develop some machinery first. A spacetime (M, g) is ***future (past) distinguishing*** if $I^+(p) = I^+(q)$ $\left(I^-(p) = I^-(q)\right)$ implies $p = q$. It is said to be ***distinguishing*** if it's either future or past distinguishing.

Proposition 6.34 *A strongly causal spacetime (M, g) is future and past distinguishing.*

Proof. Assume $I^+(p) = I^+(q)$. Assume $p \neq q$, so let U and V be disjoint open sets around p and q. Let $r \in I^+(p) \cap U$. Then $r \in I^+(q)$, so pick $s \in V$ such that $s \in I^+(q) \cap I^-(r)$. Then $s \in I^+(p)$. Thus there is a trip from p to r to s which must intersect U in a disconnected set, hence M is not strongly causal. The arguments works if we assume $I^-(p) = I^-(q)$. $\square$

Corollary 6.35 *A globally hyperbolic spacetime (M, g) is future and past distinguishing.*

The fact that a globally hyperbolic spacetime is future and past distinguishing and the sets $J^{\pm}(p)$ are closed leads to useful properties relating past and future sets.

Proposition 6.36 *If (M, g) is globally hyperbolic, then for all $p, q \in M$ we have*

(1) *$I^+(q) \subset I^+(p)$ if and only if $I^-(p) \subset I^-(q)$.*

(2) *$J^+(q) \subset J^+(p)$ if and only if $J^-(p) \subset J^-(q)$.*

Remark: The only property we use is that the sets $J^{\pm}(p)$ are closed for all $p \in M$.

Proof. Both (1) and (2) rely on the following easily established characterizations

$$J^+(p) = \overline{J^+(p)} = \{q \in M : I^+(q) \subset I^+(p)\}$$

$$J^-(q) = \overline{J^-(q)} = \{p \in M : I^-(p) \subset I^-(q)\}.$$

For (1), assume $I^+(q) \subset I^+(p)$. By the above characterization $q \in J^+(p)$. Therefore $p \in J^-(q)$ so $I^-(p) \subset I^-(q)$. The converse is similar.

For (2), assume $J^+(q) \subset J^+(p)$. Since $q \in J^+(p)$, we have $I^+(q) \subset I^+(p)$. By (1) we have $I^-(p) \subset I^-(q)$. Therefore for any $r \in J^-(p)$ we have $I^-(r) \subset I^-(p) \subset I^-(q)$ which implies $r \in J^-(q)$. The converse is similar. □

Now put a measure μ on M such that $\mu(U) > 0$ for any open set $U \neq \emptyset$ and $\mu(M) = 1$. One can construct a measure with these properties using a partition of unity. We define the ***future and past volumes*** $t^{\pm}$ as $t^+(p) = \mu\big(J^+(p)\big)$ and $t^-(p) = \mu\big(J^-(p)\big)$, respectively.

Proposition 6.37 *If (M, g) is globally hyperbolic, then t^- is bounded, strictly increasing along every future directed causal curve γ, and continuous along γ. Analogous statements hold for t^+.*

Proof. t^- is bounded since $\mu(M) = 1$. Let γ be a future directed causal curve and fix $q, p \in \gamma$ with $q \in J^+(p)$ and $q \neq p$. By (2) in proposition 6.36, we have $J^-(p) \subset J^-(q)$ which implies $t^-(q) \geq t^-(p)$. Moreover, since $q \neq p$ we have $I^-(q) \neq I^-(p)$ since (M, g) is past distinguishing. But $I^-(p) \subset I^-(q)$ since $p \in J^-(q)$. Thus we can find an open set in $I^-(q) \setminus I^-(p)$ which implies $t^-(q) > t^-(p)$. Therefore t^- is strictly increasing along γ.

To show t^- is continuous along γ let $\{p_n\}$ be a sequence of points on γ converging to $p \in \gamma$. If γ were not continuous at p, then we could find an $\epsilon > 0$ and a subsequence $\{p_{n'}\}$ which satisfies $|t^-(p) - t^-(p_n)| \geq \epsilon$. This forces $\mu(M) = \infty$. □

Now we are ready to prove the converse of Theorem 6.33.

Theorem 6.38 *If (M, g) is globally hyperbolic, then M admits a Cauchy surface S and M is topologically $\mathbb{R} \times S$.*

Proof. Let t^- and t^+ be future and past volumes on (M, g). Let γ be a future directed causal curve. Consider the continuous function $f = t^-/t^+$ on M. The sets $S_r = \{p : f(p) = r \in \mathbb{R}\}$. Each S_r is closed (because f is continuous) and achronal (because f is strictly increasing along any future directed causal curve).

Now we show every inextendible causal curve intersects every S_r. This must occur if t^- goes to zero along every past inextendible causal curve and if t^+ goes to zero along

every future inextendible causal curve, since then f will attain all values in $(0, \infty)$. Let γ be any past inextendible causal curve starting at q such that t^- does not approach zero along γ. Then there must exist a point $p \in I^-(r)$ for every $r \in \gamma$. Therefore γ is contained in the compact set $J^+(p) \cap J^-(q)$. By proposition 6.18, γ must have an endpoint in $J^+(p) \cap J^-(q)$ which contradicts γ being past inextendible. Similarly, t^+ must approach zero along every future inextendible causal curve.

By proposition 6.31 S_r and $S_{r'}$ are homeommorphic for all $r, r' \in \mathbb{R}$. Call $S = S_1$ and let $\psi : S_r \to S$ be a homeomorphism. Each $p \in M$ lies on some S_r, so we define $\xi : M \to S$ by declaring $\xi(p) = \psi_r(p)$. Therefore the desired homeomorphism $\phi : M \to (0, \infty) \times S$ is given by $\phi(p) = \big(f(p), \xi(p)\big)$. $\square$

Thus any nontrivial topology in a globally hyperbolic spacetime must reside in its Cauchy surface. We summarize the results of this section. By combining proposition theorem 6.33 and theorem 6.38 we have the following fundamental result of globally hyperbolic spacetimes:

Theorem 6.39 *A spacetime (M, g) is globally hyperbolic if either of the following hold:*
(1) *M admits a Cauchy surface S.*
(2) *M is strongly causal and $J^+(p) \cap J^-(q)$ is compact for any $p, q \in M$.*
Moreover, M is topologically $\mathbb{R} \times S$.

6.5 Lorentzian Distance Function

Many results in Riemannian geometry are established by using the Riemannian distance function. Indeed the celebrated Hopf-Rinow Theorem gives certain equivalent conditions of completeness which establish when any two points on a Riemannian manifold can be joined by a distance-minimizing geodesic. We seek an analogous result for the proper time of timelike paths in spacetimes. Unlike Riemannian geometry, Theorem 3.7 shows that proper time is locally maximized precisely by timelike geodesics. Therefore we expect that, after finding certain conditions on the spacetime, timelike geodesics will maximize proper time. The role that completeness plays in Riemannian geometry will be replaced with global hyperbolicity.

Let (M, g) be a spacetime. Let $\Omega_{p,q}$ denote the collection of continuous future directed causal curves from p to q. Recall from the argument given above lemma 6.7 that *continuous* future directed causal curves satisfy a local Liptschitz condition which implies that they are differentiable almost everywhere. If $\gamma : [a, b] \to M$ is such a curve, then we define the length of γ as

$$L(\gamma) = \int_a^b \sqrt{-g\big(\gamma'(s), \gamma'(s)\big)}ds.$$

If γ is a timelike curve then $L(\gamma) = c\tau$ where τ is the proper time of γ. We define the ***Lorentzian distance function*** $d : M \times M \to [0, \infty]$ as

$$d(p, q) = \begin{cases} \sup\{L(\gamma) : \gamma \in \Omega_{p,q}\}, & \text{if } q \in J^+(p) \\ 0, & \text{otherwise} \end{cases}$$

The reason for taking the supremum as oppose to the infimum relates back to the fact that timelike geodesics locally *maximize* proper time. In fact for $q \in J^+(p)$, $\inf\{L(\gamma) :$

$\gamma \in \Omega_{p,q}\} = 0$ since we can always approximate timelike paths by piecewise null curves.

Another peculiar feature of the Lorentzian distance function is that there's no reason it can't take on value ∞. This is easily seen with spacetimes with topology $S^1 \times \mathbb{R}^3$ with the induced metric from Minkowski space.

Lemma 6.40 *The Lorentzian distance function obeys a reverse triangle inequality. More precisely, if $r \in J^+(q)$ and $q \in J^+(p)$, then $d(p,r) \geq d(p,q) + d(q,r)$.*

Proof. Let γ be a path from p to q and λ a path from q to r. Then we have

$$d(p,r) \geq L(\gamma) + L(\lambda).$$

Taking the supremum over all $\gamma \in \Omega_{p,r}$ yields $d(p,r) \geq d(p,q) + L(\lambda)$. Then taking the supremum over all $\lambda \in \Omega_{r,q}$ yields $d(p,r) \geq d(p,q) + d(q,r)$. □

Proposition 6.41 *The Lorentzian distance function is lower semi-continuous wherever it's finite. Also, if $d(p,q) = \infty$, $p_n \to p$, and $q_n \to q$, then $\lim_{n\to\infty} d(p_n, q_n) = \infty$.*

Proof. Fix $p, q \in M$ and $\epsilon > 0$. We seek neighborhoods U and V of p and q, respectively, such that for all $r \in U$ and $s \in V$, $d(p,q) < d(r,s) + \epsilon$.

If $d(p,q) = 0$, then we're done. Let's first assume $0 < d(p,q) < \infty$. We can find a future directed timelike curve γ from p to q such that $d(p,q) = L(\gamma) + \epsilon/3$. Let U and V be convex normal neighborhoods about p and q, respectively, such that the lorentzian distance between any two points in U and V is less than $\epsilon/3$. Choose $p' \in \gamma \cap U$ and $q \in \gamma \cap V$ and let $U' = I^-(p') \cap U$ and $V' = I^+(q') \cap V$. Fix $r \in U'$ and $s \in V'$. Let λ be

the curve which starts at r travels to p' along a timelike geodesic, travels along γ to q', and then travels to r' along a timelike geodesic. Then we have $L(\lambda) > L(\gamma) - 2\epsilon/3 = d(p,q) - \epsilon$. Thus $d(r,s) \geq L(\lambda) > d(p,q) - \epsilon$.

Now suppose $d(p,q) = \infty$ with $p_n \to p$, $q_n \to q$, and $\liminf d(p_n, q_n) = R < \infty$. Fix $\epsilon > 0$. Since $d(p,q) = \infty$ there exists a future directed timelike curve γ from p to q with length $L(\gamma) > R + 2\epsilon$. Let U and V be convex normal neighborhoods around p and q, respectively, such that the Lorentzian distance between any two points in U is and V is less than ϵ. Define U' and V' like above so that for any $r \in U'$ and $s \in V'$, we can find a curve λ such that $L(\lambda) > L(\gamma) - 2\epsilon$. Therefore $d(r,s) \geq L(\lambda) > L(\gamma) - 2\epsilon > R$. Since there exist arbitrarily small neighborhoods U' and V' with this property, we contradict $\liminf d(p_n, q_n) = R$. □

We will see that d is in fact continuous and finite for globally hyperbolic spacetimes. However, we need to introduce some new terminology and a few lemmas before we get there. Let γ and $\{\gamma_n\}$ be continuous causal curves defined each defined on the closed interval $[a,b]$. The sequence $\{\gamma_n\}$ is said to ***converge to γ in the C^0 topology on curves*** if $\gamma_n(a) \to \gamma(a)$, $\gamma_n(b) \to \gamma(b)$, and given any open set U containing γ, there is an integer N such that $\gamma_n \subset U$ for all $n \geq N$. For strongly causal spacetimes, the following lemma gives an important relationship between limit curves and curves which converge in the C^0 topology on curves.

Lemma 6.42 *Let (M,g) be a strongly causal spacetime. Suppose that $\{\gamma_n\}$ is a sequence future (past) directed causal curves defined on $[a,b]$ such that $\gamma_n(a) \to p$ and $\gamma_n(b) \to q$.*

If $\gamma : [a, b] \to M$ is a future (past) directed causal curve with $\gamma(a) = p$ and $\gamma(b) = q$ and a limit curve of $\{\gamma_n\}$, then there is a subsequence of $\{\gamma_{n'}\}$ of $\{\gamma_n\}$ which converges to γ in the C^0 topology on curves.

Proof. Let U be an open set with $\gamma \subset U$. Cover the compact image of γ with local causality neighborhoods $\{V_1, \ldots, V_k\}$ such that each $V_i \subset U$. We can find a subdivision of $a = s_0 < s_1 < \cdots < s_j = b$ of $[a, b]$ such that for all $0 \leq i \leq j-1$, each pair $\gamma(s_i), \gamma(s_{i+1})$ lies in V_h for some h. Let $\{\gamma_{n'}\}$ be the subsequence of $\{\gamma_n\}$ from the definition of limit curve. For each n', put $p(0, n') = \gamma_{n'}(a)$, $p(j, n') = \gamma_{n'}(b)$, and $p(i, n') \in \gamma_{n'}$ such that the sequence $\{p(i, n')\}_{n'}$ converges to $\gamma(s_i)$. Since $\gamma(s_{i+1})$ lies in the causal future of $\gamma(s_i)$, strong causality implies there exists an integer N_1 such that $p(i+1, n') \in J^+\big(p(i, n')\big)$ for all $n' \geq N_1$. Also, there is some N_2 such that $p(i, n')$ and $p(i+1, n')$ lie in V_h for all $n' \geq N_2$. Thus for $n' \geq \max\{N_1, N_2\}$, the portion of $\gamma_{n'}$ joining $p(i, n')$ to $p(i+1, n')$ must lie entirely in V_h otherwise $\gamma_{n'}$ would have to leave and reenter V_h. Therefore $\gamma_{n'} \subset V_1 \cup \cdots \cup V_k \subset U$ for all $n' \geq N$. □

In fact, the converse of Lemma 6.42 is also true but it's harder to prove and not necessary for our purposes. The next proposition shows the length function is upper semi-continuous with respect to the C^0 topology on curves.

Proposition 6.43 *If (M, g) is strongly causal and $\{\gamma_n\}$ are continuous causal curves which converge to the continuous causal curve $\gamma \in \Omega_{p,q}$ in the C^0 topology on curves, then $L(\gamma) \geq \limsup L(\gamma_n)$.*

Proof. We want to show that for any $a \in \mathbb{R}$ such that $L(\gamma) < a$, we can find a neighborhood U containing γ such that any causal curve λ contained in U also satisfies $L(\lambda) < a$. Let $\xi = \{p_0, p_1, \ldots, p_k\}$ denote a finite sequence of points along γ, beginning at $p_0 = p$ and ending at $p_k = q$, such that any consecutive pair p_i, p_{i+1} are contained in a local causality neighborhood U_i which also contains the portion of γ from p_i to p_{i+1}. Let γ_i denote the unique geodesic segment from p_i to p_{i+1} and put $\gamma_\xi = \bigcup_{i=0}^{k-1} \gamma_i$. By theorem 3.7 $L(\gamma) < L(\gamma_\xi)$. Choose ξ such that $L(\gamma_\xi) < L(\gamma)/2 + a/2$ and U_i such that U_i only intersects U_{i-1}, U_i, and U_{i+1}. Since the length of a geodesic is a continuous function on its endpoints in a convex normal neighborhood, we can choose local causality neighborhoods $\{V_0, V_1, \ldots, V_k\}$ small enough such that $p_i \in V_i$ and any causal geodesic from a point of V_i to a point of V_{i+1} must differ in length from $L(\gamma_i)$ by less than $\big[a - L(\gamma)\big]/2k$. Define

$$W_i = \bigcup \big\{ \langle r, s \rangle : r \in V_i \text{ and } s \in V_{i+1} \big\}.$$

We have $W_i \subset U_i$ since U_i is a local causality neighborhood. So just like U_i, W_i intersects W_j only if $j = i - 1, i, i + 1$. Let $U = \bigcup_i W_i$ and fix $\lambda \subset U$ to be a future directed causal curve. λ passes through the W_i consecutively. Moreover λ must pass through each V_i since λ meets $W_{i-1} \cap W_i$. Therefore λ' contains the set of points $\xi' = \{p'_0, p'_1, \ldots, p'_k\}$ with $p'_i \in V_i$. Let λ_i be the unique geodesic which connects p'_i to p'_{i+1} and $\lambda_{\xi'} = \bigcup_i \lambda_i$.

Therefore, just like above, theorem 3.7 implies

$$\begin{aligned} L(\lambda) &< L(\lambda_{\xi'}) \\ &< L(\gamma_\xi) + k\left(\frac{a - L(\gamma)}{2k}\right) \\ &< \frac{L(\gamma)}{2} + \frac{a}{2} + k\left(\frac{a - L(\gamma)}{2k}\right) \\ &= a \end{aligned}$$

which is what we wanted to show. □

Proposition 6.44 *If (M, g) is globally hyperbolic, then the Lorentzian distance function d is finite and continuous on $M \times M$.*

Proof. Fix $q \in J^+(p)$. To prove that $d(p, q)$ is finite, cover the compact set $J^+(p) \cap J^-(q)$ with a finite number of local causality neighborhoods $\{U_1, \ldots, U_m\}$ such that every causal curve in each U_i has length at most one. Any causal curve γ from p to q can only enter each U_i once so $L(\gamma) \leq M$. Hence $d(p, q) \leq m$.

From proposition 6.41 we know d is lower semi-continuous. Therefore it suffices to show d is upper semi-continuous. Assume otherwise. Then we could find a $\delta > 0$ and sequences $\{p_n\}$ and $\{q_n\}$ converging to p and q respectively, such that $d(p_n, q_n) \geq d(p, q) + 2\delta$ for all n. By definition of $d(p_n, q_n)$, we can find a future directed causal curve γ_n from p_n to q_n with $L(\gamma_n) \geq d(p, q) + \delta$ for each n. Using the same technique we used in the proof of theorem 6.33, we can find a future directed causal curve γ from p to q which is a limit curve of $\{\gamma_n\}$. By lemma 6.42, a subsequence $\{\gamma_{n'}\}$ of $\{\gamma_n\}$ converges to γ in the C^0 topology on curves. By proposition 6.43, $L(\gamma) \geq \limsup L(\gamma_n) \geq d(p, q) + \delta$.

But this contradicts the definition of $d(p,q)$. $\square$

For any points $p, q \in M$ with $q \in J^+(p), q \neq p$, the curve $\gamma \in \Omega_{p,q}$ is said to be ***maximal*** if $L(\gamma) = d(p,q)$. If γ is inextendible in any way, then we'll say γ is maximal if it's maximal between any two of its points.

Notice that if $\gamma : [a,b] \to M$ in $\Omega_{p,q}$ is maximal, then lemma 6.40 implies that for all s, t with $a \leq s \leq s' \leq b$, we have $d\big(\gamma(s), \gamma(s')\big) = L\big(\gamma|_{[s,s']}\big)$.

Lemma 6.45 *If $\gamma \in \Omega_{p,q}$ is maximal, then γ is a geodesic.*

Proof. If $q \notin I^+(p)$, then γ is a null geodesic by proposition 6.1. Now suppose $q \in I^+(p)$ and γ from p to q is maximal. For any $\gamma(s) \in \gamma$, we can find a $\delta > 0$ such that a convex normal neighborhood contains $\gamma\big([s-\delta, s+\delta]\big)$. By the comment below the definition of maximal, $\gamma|_{[s-\delta,s+\delta]}$ is maximal, so theorem 3.7 implies $\gamma|_{[s-\delta,s+\delta]}$ is a geodesic. The result follows since $\gamma(s) \in \gamma$ was arbitrary. $\square$

Now we can prove the main result of this section. Between any two causally separated points in a globally hyperbolic spacetime, there exists a geodesic γ which connects them. Theorem 6.46 is the analogue of the Hopf-Rinow theorem in Riemannian geometry, and just like in the case of Riemannian geometry, the geodesic may not be unique.

Theorem 6.46 *Let (M,g) be globally hyperbolic. Then given any $p, q \in M$ with $q \in J^+(p)$, there is a maximal geodesic $\gamma \in \Omega_{p,q}$.*

Proof. If $q \notin I^+(p)$, then by proposition 6.1 γ is a null geodesic. γ is maximal because any other curve connecting p to q must also be a null geodesic which has length zero. Consider $q \in I^+(p)$. Let f be the continuous function used in the proof of Theorem 6.38. Recall that f is strictly increasing along any future directed causal curve. Choose $s_0 \in \mathbb{R}$ such that $f(p) < s_0 < f(q)$. Then $K = J^+(p) \cap J^-(q) \cap f^{-1}\big(\{s_0\}\big)$ is compact. Moreover $f^{-1}\big(\{s_0\}\big)$ is a Cauchy surface, that intersects $J^+(p) \cap J^-(q)$. Therefore every causal curve in $\Omega_{p,q}$ must intersect K. For each positive integer n, we can find curves $\gamma_n \in \Omega_{p,q}$ such that

$$d(p,q) \geq L(\gamma_n) \geq d(p,q) - \frac{1}{n}.$$

Let $r_n \in \gamma_n \cap K$. Since K is compact, a subsequence $\{r_{n'}\}$ converges to $r \in K$. Using techniques similar in the beginning of the proof of theorem 6.33, there is a limit curve $\gamma \in \Omega_{p,q}$ of the sequence $\{\gamma_{n'}\}$ passing through r. By lemma 6.42, a subsequence $\{\gamma_{n''}\}$ converges to γ in the C^0 topology on curves. By proposition 6.43, we have

$$L(\gamma) \geq \limsup L(\gamma_{n''}) \geq d(p,q)$$

Thus $L(\gamma) = d(p,q)$ and so it's a maximal curve. By lemma 6.45, it's a geodesic. $\square$

7 Singularity Theorems

Intuitively, a singularity is a place in a physical theory where something goes bad. For example, the Coulomb solution of a point charge at the origin has an infinite charge density at $r = 0$. Similarly in Newtonian gravity, if one considers a spherical, nonrotating

shell of dust released from rest, then there will be an infinite mass density as all the dust simultaneously reach the origin. Likewise, the Schwarzschild solution and FRW solutions admit timelike paths which observers can end their existence or begin their existence. Moreover, observers following these paths will feel a lot of discomfort since these paths have a curvature singularity. Historically, the curvature singularities were not as troubling to physicists as observers beginning or ceasing their existence. This is because a curvature singularity is thought to arise from the stress energy tensor and hence would be a result of an infinite density of mass or energy much like the Coulomb and Newtonian example above. These infinities are merely a lack of understanding of the internal structure of matter, which one would need a full understanding of quantum gravity to resolve. However, it was thought that beginning or ceasing one's existence was physically impossible. Thus it was hypothesized that this pathological behavior which occurs in the Schwarzschild solution and the FRW solutions are due to the high degrees of symmetry in the solutions. For example, if the spherical shell of dust in Newtonian gravity is perturbed, then the infinite mass singularity would not occur. However, in 1965 Roger Penrose showed the situation in general relativity is quite different. Under mild assumptions, he showed that spacetimes modeling gravitational collapse will always contain at least one null geodesic with finite affine parameter.

This has become the standard criterion for a singularity in general relativity. More concretely, a spacetime (M, g) is ***singular*** or ***contains a singularity*** if there exist timelike or null geodesics with affine parameters which don't take on all values of $\mathbb{R}$. Just like in the Riemannian case, these geodesics are said to be ***incomplete***. Of course one can artificially remove points from a nonsingular spacetime thus making it singular.

To avoid this scenario, we restrict our attention to ***inextendible spacetimes***, i.e., spacetimes which are not isometric to a proper subset of another spacetime. It should also be noted that a singular spacetime does not imply a curvature singularity which can be seen by considering conical spacetimes. Whether or not curvature singularities are "generic" is still an open question.

The singularity theorems prove existence of an incomplete geodesic by contradicting the existence of maximal timelike geodesics and or when a null geodesic remains on the boundary of the future of a point. Thus we begin by establishing when geodesics fail to maximize the proper time between points and when a null geodesic fails to remain on the boundary of the future of a point.

7.1 Timelike Congruences and Energy Conditions

Let $U \subset M$ be open. A ***congruence of curves*** in U is a three-parameter family of curves such that there is a unique curve of the family passing through each $p \in U$. The vector field formed by the vectors tangent to the curves is called the ***tangent vector field***. A congruence is ***timelike*** or ***null***) if the tangent vector field is timelike or null, respectively. We will only consider timelike congruences in this section.

Let u be a tangent vector field of a timelike congruence such that $g(u,u) = -c^2$, i.e. u is the four-velocity of the family of curves. For any of the curves, we can define an affine parameter τ by $u(\tau) = u^a \nabla_a \tau = 1$. This is equivalent to the usual definition $\tau = \frac{1}{c}\int \sqrt{-g(u,u)}ds$, however the definition $u(\tau) = 1$ generalizes for null congruences.

Conversely, any unit timelike vector field defines locally a timelike congruence by solving the differential equations $dx^a/d\tau = u^a(x)$ in any local chart. A prior the solution depends on four constants, but one of them can be absorbed into the affine parameter τ. Thus, prescribing a unit timelike tangent vector field u gives rise a timelike congruence.

We define the spatial metric h of the congruence by

$$h_{ab} = g_{ab} + c^{-2}u_a u_b. \tag{7.1}$$

Therefore $h^a{}_b = g^{ac}h_{bc}$ is the projection operator onto the subspace of the tangent space perpendicular to u. The ***acceleration*** of the timelike congruence is the vector field defined by $a = \nabla_u u$. Notice that $g(a,u) = 0$ so a is spacelike. Using the projection operator, we can decompose the $(0,2)$ tensor, ∇u, with components $\nabla_b u_a$, into the following:

$$\begin{aligned}
\nabla_b u_a &= g^c{}_a g^d{}_b \nabla_d u_c \\
&= [-c^{-2}u^c u_a + h^c{}_a][-c^{-2}u^d u_b + h^d{}_b]\nabla_d u_c \\
&= c^{-4}u^c u_a u^d u_b \nabla_d u_c - c^{-2}(u^c u_a h^d{}_b + u^d u_b h^c{}_a)\nabla_d u_c + h^c{}_a h^d{}_b \nabla_d u_c. \qquad (7.2)
\end{aligned}$$

Let's look at the first, second, and third terms of eq. (7.1). The first term is $c^{-4}u^c u_a u^d u_b \nabla_d u_c = c^{-4}u_a u_b g(\nabla_u u, u) = c^{-4}u_a u_b g(a,u) = 0$. Using eq. (7.2) the second term satisfies

$$\begin{aligned}
-c^{-2}u^c u_a h^d{}_b \nabla_d u_c &= -c^{-2}u^c u_a (g^d{}_b + c^{-2}u^d u_b)\nabla_d u_c \\
&= -c^{-2}u^c u_a \nabla_b u_c - c^{-4}u^c u_a u^d u_b \nabla_d u_c \\
&= -c^{-2}u_a \nabla_b g(u,u) - c^{-4}u_a u_b g(\nabla_u u, u) \\
&= -c^{-2}u_a \nabla_b (-c^2) - c^{-4}u_a u_b g(a,u) \\
&= 0.
\end{aligned}$$

However the third term is nonzero

$$\begin{aligned} -c^{-2}u^d u_b h^c{}_a \nabla_d u_c &= -c^{-2}u^d u_b (g^c{}_a + c^{-2}u^c u_a)\nabla_d u_c \\ &= -c^{-2}u^d u_b g^c{}_a \nabla_d u_c \\ &= -c^{-2}u_b (\nabla_u u)_a \\ &= -c^{-2}u_b a_a. \end{aligned}$$

Therefore we have the following formula for ∇u

$$\nabla_b u_a = -c^{-2}u_b a_a + h^c{}_a h^d{}_b \nabla_d u_c.$$

Now we introduce the function θ and (0,2) tensors σ and ω which are defined by

$$\begin{aligned} \theta &= \nabla_a u^a \\ \sigma_{ab} &= \frac{1}{2}h^c{}_a h^d{}_b (\nabla_d u_c + \nabla_c u_d) - \frac{\theta}{3}h_{ab} \\ \omega_{ab} &= \frac{1}{2}h^c{}_a h^d{}_b (\nabla_d u_c - \nabla_c u_d). \end{aligned}$$

Thus we can write $\nabla_b u_a$ as

$$\begin{aligned} \nabla_b u_a &= -c^{-2}u_a u_b + h^c{}_a h^d{}_b \nabla_d u_c \\ &= -c^{-2}u_b a_a + \frac{\theta}{3}h_{ab} + \sigma_{ab} + \omega_{ab}. \end{aligned} \tag{7.3}$$

θ, σ, and ω are known as the ***expansion***, ***shear tensor***, and ***rotation tensor***, respectively. Physically, these quantities represent the expansion, deformation, and twist of a small volume element along the curves of the congruence which justifies their names.

Let us quickly remark that u is integrable (i.e. proportional to a graident: $u_a = f\nabla_a s$ for some functions f and s) if and only if $\omega = 0$. This follows from the torsion-free property of the connection. In this case, the congruence is said to be ***irrotational***.

Now we embark on deriving the fundamental Raychaudhuri's equation. This is the key equation used in the proof of the singularity theorems. Since it's so crucial, we give a step-by-step derivation of it. Let us start with

$$\begin{aligned} u^c\nabla_c\nabla_b u^a &= u^c(\nabla_b\nabla_c u^a + R^a{}_{dcb}u^d) \\ &= u^c\nabla_b\nabla_c u^a - R^a{}_{dbc}u^c u^d \\ &= \nabla_b(u^c\nabla_c u^a) - (\nabla_b u^c)(\nabla_c u^a) - R^a{}_{dbc}u^c u^d. \end{aligned} \tag{7.4}$$

Tracing over components a and b, we have

$$\begin{aligned} \frac{d\theta}{d\tau} &= u(\theta) \\ &= u^c\nabla_c\nabla_b u^b \\ &= \nabla_b a^b - (\nabla_b u^c)(\nabla_c u^b) - R_{cd}u^c u^d \end{aligned} \tag{7.5}$$

Let us focus on the middle term, $(\nabla_b u^c)(\nabla_c u^b)$. Using eq. (7.3) we have

$$\begin{aligned} (\nabla_b u^c)(\nabla_c u^b) &= (\nabla_b u_a)(\nabla^a u^b) \\ &= \left(-c^{-2}u_b a_a + \frac{\theta}{3}h_{ab} + \sigma_{ab} + \omega_{ab}\right)\left(-c^{-2}u^a a^b + \frac{\theta}{3}h^{ba} + \sigma^{ba} + \omega^{ba}\right) \end{aligned} \tag{7.6}$$

To evaluate eq. (7.6), let's consider each term separately. $u_b a_a u^a a^b = g(u,a)^2 = 0$. Also, $u_b h^{ba} = u_b(g^{ba} + c^{-2}u^b u^a) = u_a - u_a = 0$. Therefore $u_b\sigma^{ba} = u_b\omega^{ba} = 0$ by definition of the shear and rotation tensor. Notice we could have expected this since h_{ab}, σ_{ab}, and ω_{ab} are "purely spatial" tensors. Now we evaluate

$$\frac{\theta}{3}h_{ab}\frac{\theta}{3}h^{ab} = \frac{\theta^2}{3}$$

Since h_{ab} is symmetric, we have

$$\begin{aligned}
h_{ab}(\sigma^{ba} + \omega^{ba}) &= h^{ab}(\sigma_{ab} + \omega_{ab}) \\
&= h^{ab}(h^{c}{}_{a} h^{d}{}_{b} \nabla_d u_c) - \frac{\theta}{3} h_{ab} h^{ab} \\
&= (g^{ab} + c^{-2} u^a u^b)(h^{c}{}_{a} h^{d}{}_{b} \nabla_d u_c) - \theta \\
&= h^{cb} h^{d}{}_{b} \nabla_d u_c - \theta \\
&= (g^{cb} + c^{-2} u^c u^b) h^{d}{}_{b} \nabla_d u_c - \theta \\
&= h^{dc} \nabla_d u_c - \theta \\
&= (g^{dc} + c^{-2} u^d u^c) \nabla_d u_c - \theta \\
&= \nabla_d u^d + c^{-2} \nabla_u (-c^2) - \theta \\
&= \theta - \theta \\
&= 0.
\end{aligned}$$

Lastly, we calculate

$$\begin{aligned}
\sigma^{ab}\omega_{ba} &= \left[\frac{1}{2}h^{c}{}_{a}h^{d}{}_{b}(\nabla_d u_c + \nabla_c u_d) - \frac{\theta}{3}h_{ab}\right]\left[\frac{1}{2}h^{ca}h^{db}(\nabla_d u_c - \nabla_c u_d)\right] \\
&= \left[\frac{1}{2}h_{ca}h_{db}(\nabla^d u^c + \nabla^c u^d) - \frac{\theta}{3}h_{ab}\right]\left[\frac{1}{2}h^{ca}h^{db}(\nabla_d u_c - \nabla_c u_d)\right] \\
&= \frac{1}{4}(3)(3)(\nabla^d u^c + \nabla^c u^d)(\nabla_d u_c - \nabla_c u_d) - \frac{\theta}{6}h_{ab}h^{cb}h^{db}(\nabla_d u_c - \nabla_c u_d) \\
&= \frac{9}{4}(\nabla^d u^c \nabla_d u_c - \nabla^d u^c \nabla_c u_d + \nabla^c u^d \nabla_d u_c - \nabla^c u^d \nabla_c u_d) - \frac{\theta}{6}h_{ab}h^{cb}h^{db}(\nabla_d u_c - \nabla_c u_d) \\
&= \frac{9}{4}(\nabla^d \nabla_d(-c^2) - \nabla_u \theta + \nabla_u \theta - \nabla^c \nabla_c(-c^2) - \frac{\theta}{6}h_{ab}h^{cb}h^{db}(\nabla_d u_c - \nabla_c u_d) \\
&= 0 - \frac{\theta}{6}(g_{ab} + c^{-2}u_a u_b)(g^{ca} + c^{-2}u^c u^b)h^{db}(\nabla_d u_c - \nabla_c u_d) \\
&= \frac{\theta}{6}(\delta^{c}{}_{b} + c^{-2}u^c u_b + c^{-2}u^c u_b + c^{-2}u^c u_b)h^{db}(\nabla_d u_c - \nabla_c u_d) \\
&= \frac{\theta}{6}\delta^{c}{}_{b}h^{db}(\nabla_d u_c - \nabla_c u_d) \\
&= \frac{\theta}{6}h^{db}(\nabla_d u_b - \nabla_b u_d) \\
&= 0.
\end{aligned}$$

The last equality follows since h^{db} is symmetric. Incorporating these results into eq. (7.6), we find

$$\begin{aligned}
(\nabla_b u^c)(\nabla_c u^b) &= \frac{\theta^2}{3} + \sigma_{ab}\sigma^{ba} + \omega_{ab}\omega^{ba} \\
&= \frac{\theta^2}{3} + \sigma_{ab}\sigma^{ab} - \omega_{ab}\omega^{ab}.
\end{aligned}$$

We used the symmetry, $\sigma_{ab} = \sigma_{ba}$, and antisymmetry, $\omega_{ab} = -\omega_{ba}$, in the last line. Therefore eq. (7.5) becomes

$$\frac{d\theta}{d\tau} = \nabla_b a^b + \omega_{ab}\omega^{ab} - \frac{\theta^2}{3} - \sigma_{ab}\sigma^{ab} - R_{ab}u^a u^b. \tag{7.7}$$

Eq. (7.7) is the fundamental ***Raychaudhuri equation***. We will use it to find conditions when $d\theta/d\tau \leq 0$. Why do we want this? If we can show $\theta \to -\infty$ in a finite amount of

proper time, then we are seeing either 1) a breakdown of the congruence or 2) singular behavior. Let us ask under what conditions is this true? If were considering a timelike geodesic, then $a = \nabla_u u = 0$ so the first term in eq. (7.7) vanishes. We see that $\sigma_{ab}\sigma^{ab} \geq 0$ and $\omega_{ab}\omega^{ab} \geq 0$,

$$\sigma_{ab}\sigma^{ab} = \sigma^{cd}\sigma^{ab}g_{ca}g_{db} = \sigma^{cd}\sigma^{ab}h_{ca}h_{db} \geq 0,$$

since h is a spatial metric. The same argument shows $\omega_{ab}\omega^{ab} \geq 0$. So by the Raychaudhuri equation (7.7), the shear tensor σ induces contraction while the rotation tensor ω induces expansion; this last part might be expected by analogy with the centrifugal force. Therefore ω is an undesirable term so let us ask under what conditions is $\omega = 0$? The following lemma shows this is the case when the congruence consists of timelike geodesics traversing orthogonal to a spacelike hypersurface.

Lemma 7.1 *Let Σ be a spacelike hypersurface in (M, g). There exists a timelike geodesic congruence of curves emanating from Σ orthogonally such that $\omega = 0$ (i.e. the congruence is irrotational).*

Proof. Fix $p \in \Sigma$ and let (T, X, Y, Z) be a synchronous coordinate system about p like the one constructed in theorem 3.7. Let u be the tangent vectors of the orthogonal timelike geodesics emanating through Σ in this coordinate system and such that $g(u, u) = -c^2$. In the synchronous coordinates we have $u^i = 0$ for $i = 1, 2, 3$ and $u^0 = \pm 1$. We want to show $\Omega_{ab} = \nabla_a u_b - \nabla_b u_a = 0$ since this will imply $\omega = 0$. For $i, j = 1, 2, 3$ we have $\Omega_{ij} = 0$ since $u^i = 0$. Also, since u is the tangent vector of a geodesic and $g(u, u) = -c^2$, we have $u^a\Omega_{ab} = 0$, and since $u^i = 0$ this implies $\Omega_{0i} = 0$. Therefore $\Omega_{ab} = 0$. By covering

Σ with synchronous coordinate systems, we find our desired congruence. $\square$

There's a very nice physical interpretation of lemma 7.1. If one fixes a spacelike hypersurface of material particles and "releases" them from rest, then the particles will follow timelike geodesics which are orthogonal to the spacelike hypersurface. The lemma above suggests that the material particles will not rotate which is what we would physically expect.

The congruence in lemma 7.1 has the following Raychaudhuri equation

$$\frac{d\theta}{d\tau} = -\frac{\theta^2}{3} - \sigma_{ab}\sigma^{ab} - R_{ab}u^a u^b. \tag{7.8}$$

θ^2 and $\sigma_{ab}\sigma^{ab}$ are both positive, so if we want $d\theta/d\tau < 0$ we want $R_{ab}u^a u^b \geq 0$. Recall that the Ricci tensor R_{ab} is related to the stress-energy tensor T_{ab} by Einstein's equation,

$$R_{ab} = \frac{8\pi G}{c^4}\left(T_{ab} - \frac{1}{2}g_{ab}T\right),$$

where $T = T^a_{\ a} = T_{ab}g^{ab}$. Therefore

$$R_{ab}u^a u^b = \frac{8\pi G}{c^4}\left(T_{ab}u^a u^b + \frac{c^2}{2}T\right). \tag{7.9}$$

We see that $R_{ab}u^a u^b \geq 0$ if the ***strong energy conditions holds:*** $T_{ab}\xi^a\xi^b + \frac{c^2}{2}T \geq 0$ for all timelike ξ^a satisfying $g(\xi,\xi) = -c^2$. Let's try and understand the physics of the strong energy condition. Consider a stress-energy tensor which is locally of the form

$$T_{ab} = \rho t_a t_b + P_1 x_a x_b + P_2 y_a y_b + P_3 z_a z_b. \tag{7.10}$$

where $\{t^a, x^a, y^a, z^a\}$ is a linearly independent set consisting of orthogonal eigenvectors of the linearly map $T^a_{\ b}$ where $\{x^a, y^a, z^a\}$ are othonormal and $g(t,t) = t^a t_a = g_{ab}t^a t^b = -c^2$.

The numbers $\{\rho, P_1, P_2, P_3\}$ are the eigenvalues of $\{t^a, x^a, y^a, z^a\}$. There is no guarantee we can put the stress-energy tensor in this form, because although $T^a_{\ b}$ is symmetric ($T^a_{\ b} = T_b^{\ a}$), the metric g_{ab} is not positive definite so we can not apply the spectral theorem. Nonetheless, it is generally believed that all physical matter has a stress energy tensor of this form. Notice that this takes the form of a perfect fluid when $P_1 = P_2 = P_3$.

Let ξ^a be any timelike vector with $\xi^a\xi_a = -c^2$. There exist numbers $\{\alpha_0, \alpha_1, \alpha_2, \alpha_3\}$ such that

$$\xi^a = \alpha_0 t^a + \alpha_1 x^a + \alpha_2 x^b + \alpha_3 x^c.$$

Therefore $\xi^a\xi_a = -c^2$ implies $-c^2 = -c^2\alpha_0^2 + \alpha_1^2 + \alpha_2^2 + \alpha_3^2$. The strong energy condition implies

$$\begin{aligned} 0 &\leq T_{ab}\xi^a\xi^b + \frac{c^2}{2}T \\ &= \rho c^2(c^2\alpha_0^2) + \sum_{i=1}^{3} P_i\alpha_i^2 + \frac{c^2}{2}\left(-c^2\rho + \sum_{i=1}^{3} P_i\right). \end{aligned}$$

If we consider $\alpha_i = 0$ for $i = 1, 2, 3$, then

$$\rho c^2 + \sum_{i=1}^{3} P_i \geq 0.$$

Similarly if we take $\alpha_1 \neq 0$ but $\alpha_2 = \alpha_3 = 0$, then we find

$$\rho c^2 + P_1 \geq 0.$$

Thus the strong energy condition implies

$$\rho c^2 + \sum_{i=1}^{3} P_i \geq 0 \quad \text{and} \quad \rho c^2 + P_i \geq 0 \ \text{ for } i = 1, 2, 3.$$

Assuming $\rho > 0$ (i.e. mass is positive), the strong energy condition will be saitisfed if there do not exist negative pressures (i.e. tensions) that are comparable in magnitude to

ρc^2.

The strong energy condition is the primary energy condition used in the singularity theorems, but let us digress to discuss other important energy conditions. The ***weak energy condition*** is satisfied if for all future directed timelike ξ^a, $T_{ab}\xi^a\xi^b \geq 0$. If one considers a stress-energy tensor of the form (7.10), then the weak energy condition implies

$$\rho \geq 0 \quad \text{and} \quad \rho c^2 + P_i \geq 0 \quad \text{for } i = 1, 2, 3.$$

By continuity, the weak energy condition implies the ***null energy condition***: $T_{ab}k^a k^b \geq 0$ for all null vectors k^a. Assuming the form (7.10), the null energy condition implies

$$\rho c^2 + P_i \geq 0 \quad \text{for } i = 1, 2, 3.$$

Probably the most physically realistic energy condition is the ***dominant energy condition***. This is satisfied if for any future directed timelike vector ξ^a, the vector $-T^a_{\ b}\xi^b$ is future directed timelike or null. The vector $T^a_{\ b}\xi^b$ physically represents the energy-momentum current density of matter as seen by the observer ξ^a. Therefore the dominant energy condition says that the speed of energy flow is always less than or equal to the speed of light. If we assume the form (7.10), the dominant energy condition implies

$$\rho c^2 \geq |P_i| \quad \text{for } i = 1, 2, 3.$$

Note that the weak energy condition implies the null energy condition and the dominant energy condition implies the weak energy condition, but otherwise the energy conditions are all independent assumptions despite their suggestive names. In particular, the strong energy condition does not imply the weak energy condition. It is "stronger" in the sense that it seems more physically probable that the weak energy condition holds

(matter is nonnegative) than the strong energy condition.

Let us now return to Raychaudhuri's equation in the form of eq. (7.8). The strong energy condition gives us the following fundamental result:

Lemma 7.2 *Let u describe an irrotational timelike geodesic congruence like the one guaranteed by lemma 7.1. If $R_{ab}u^a u^b \geq 0$ (which will be the case if the strong energy condition is satisfied) and if the expansion θ takes the value $\theta_0 < 0$ at any point on a geodesic in the congruence, then θ goes to $-\infty$ along that geodesic within finite proper time $\tau \leq 3/|\theta_0|$.*

Proof. From eq. (7.8), we have $d\theta/d\tau \leq -\theta^2/3$ which implies $d(\theta^{-1})/d\tau \geq \frac{1}{3}$. So if we parametrize the geodesic in question by τ and such that $\theta(0) = \theta_0$, then we the inequality gives $\theta^{-1} \geq \theta_0^{-1} + \tau/3$. Since $\theta_0^{-1} < 0$, within proper time $\tau = 3|\theta_0^{-1}|$, θ^{-1} must reach 0. That is, θ approaches $-\infty$ within proper time $\tau \leq 3/|\theta_0|$. □

7.2 Jacobi Fields and Conjugate Points

In general lemma 7.2 only describes a singularity in our choice of congruence. For example, if one considers the congruence of geodesics which forms the set $I^-(p)$ in Minkowski space, then the expansion will approach $-\infty$ as one approaches the point p on one of the geodesics. It is when lemma 7.2 is combined with the theory of conjugate points and the existence of maximal curves in globally hyperbolic spacetimes (section 6.5), that we will see the existence of incomplete geodesics.

Let γ be a geodesic with tangent γ'. Let u^a be the components of γ', then if the vector field η solves the geodesic deviation equation (or Jacobi equation)

$$u^c\nabla_c(u^b\nabla_b\eta^a) = R^a{}_{bcd}u^c\eta^d u^b \quad \text{equivalently} \quad \nabla_{\gamma'}(\nabla_{\gamma'}\eta) = R(\gamma',\eta)\gamma',$$

then η is said to be a ***Jacobi field*** on γ. Using parallel orthonormal basis vectors along γ, we can show that the Jacobi equation reduces to solving a linear system of 2nd order ordinary differential equations. Hence a solution always exists. The points p and q are ***conjugate*** if there exists a Jacob field η which is not identically zero but vanishes at both p and q. Conjugate points and the expansion θ are related by the following proposition.

Proposition 7.3 *Consider the timelike geodesic congruence emanating from p. Let γ be one of the timelike geodesics. Then $q \in \gamma$ is conjugate to p if and only if the expansion θ of the congruence approaches $-\infty$ along γ.*

Proof. Consider an orthonormal set of spatial vectors $\{e_1, e_2, e_3\}$ which are orthogonal to $u = \gamma' = \partial/\partial\tau$ and parallelly propagated along γ. $\{\gamma', e_1, e_2, e_3\}$ provides a basis for the tangent space at each point on γ, so the components we will work in will be with respect to this basis. Notice that $u^0 = 1$ and $u^i = 0$ for $i = 1, 2, 3$. Let η be nontrivial a Jacobi field on γ which is orthogonal to γ' and satisfies $[\gamma', \eta] = 0$. The existence of such a Jacobi field can be seen from section 4.2. Then we see that $\eta^0 = 0$. Moreover, since we're using an orthonormal basis along γ, the geodesic deviation equation becomes

$$\frac{d^2\eta^a}{d\tau^2} = R^a{}_{bcd}u^c\eta^d u^b.$$

Therefore $\eta^a(\tau)$ depends linearly on the initial data $\eta^a(0)$ and $\frac{d\eta^a}{d\tau}(0)$. Since the geodesic congruence begins at p, we must have $\eta^a(0) = 0$, so there exists a matrix $A(\tau)$ with components $A^a{}_b(\tau)$ such that

$$\eta^a(\tau) = A^a{}_b(\tau)\frac{d\eta^b}{d\tau}(0).$$

We must have $A^a{}_b(0) = 0$ and $\frac{A^a{}_b}{d\tau}(0) = \delta^a{}_b$. By definition of our congruence, η vanishes at p. Therefore q will be conjugate to p if and only if η vanishes at q if and only if A is singular at q. Therefore a necessary and sufficient condition for q to be conjugate to p is $\det(A) = 0$ at q. Now since $[u, \eta] = 0$, we have

$$\frac{d\eta^a}{d\tau} = u^b\nabla_b\eta^a = \eta^b\nabla_b u^a.$$

Now using $\eta^a(\tau) = A^a{}_b(\tau)\frac{d\eta^b}{d\tau}(0)$, we have

$$\eta^b\nabla_b u^a = \left(A^b{}_c(\tau)\frac{d\eta^c}{d\tau}(0)\right)\nabla_b u^a = \frac{dA^a{}_b}{d\tau}(\tau)\frac{d\eta^b}{d\tau}(0).$$

Thus we have $\frac{dA^a{}_b}{d\tau} = A^b{}_c\nabla_b u^a$. If we let B denote the matrix with components $B^a{}_b = \nabla_b u^a$. Then in matrix notation, we have $\frac{dA}{d\tau} = BA$. A will be nonsingular between conjugate points, so at these points we can write $B = \frac{dA}{d\tau}A^{-1}$. At these points, we have

$$\theta = B^a{}_a = \text{tr } B = \text{tr } \left(\frac{dA}{d\tau}A^{-1}\right) = \frac{1}{\det A}\frac{d}{d\tau}(\det A) = \frac{d}{d\tau}\ln|\det A|.$$

Therefore $\theta \to -\infty$ if and only if $\det A = 0$ if and only if q is conjugate to p. □

Recall from Lemma 3.4 that the timelike geodesics through p are orthogonal to spacelike hypersurfaces. So in connection with lemma 7.1 and lemma 7.2, we have

Proposition 7.4 *Let u describe the irrotational timelike geodesic congruence emanating from p. If $R_{ab}u^a u^b \geq 0$ (which will be the case if the strong energy condition is satisfied)*

and if the expansion θ takes the value $\theta_0 < 0$ at some point $r \in \gamma$ in the congruence, then within proper time $\tau \leq 3/|\theta_0|$ from r there exists a point $q \in \gamma$ conjugate to p, provided that γ can be extended that far.

In Riemannian geometry, conjugate points mark the end when a geodesic locally minimizes its length. We will show that in Lorentzian geometry, conjugate points of a timelike geodesic mark the end when the geodesic locally maximizes its proper time. Just like in Riemannian geometry, this is done by deriving the first and second variational formulas for timelike curves.

7.3 Timelike Variations Applied to Conjugate Points

In Riemannian geometry, conjugate points mark the end of minimizing geodesics. In Lorentzian geometry, conjugate points mark the end of timelike geodesics maximizing proper time. In this section we will show this using the first and second variational formulas of the length functional (which is just c times the proper time). The ideas and proofs in this section are analogous to those in the Riemannian setting.

Let $\gamma : [a, b] \to M$ be a smooth curve. A ***smooth variation*** of γ is a smooth mapping $\alpha : [a, b] \times (-\epsilon, \epsilon) \to M$, for some $\epsilon > 0$, with $\alpha(s, 0) = \gamma(s)$ for all $s \in [a, b]$. α is ***proper*** if $\alpha(a, r) = \gamma(a)$ and $\alpha(b, r) = \gamma(b)$ for all $r \in (-\epsilon, \epsilon)$. α is a ***piecewise smooth variation*** of a piecewise smooth curve γ if α is continuous and there exists a finite partition of $a = s_0 < s_1 < \cdots < s_{k-1} < s_k = b$ of $[a, b]$ such that $\gamma|_{[s_i, s_{i+1}]}$ is smooth and $\alpha|_{[s_i, s_{i+1}] \times (-\epsilon, \epsilon)}$ is a smooth variation of $\gamma|_{[s_i, s_{i+1}]}$ for each $i = 0, 1, \cdots, k - 1$.

We define the curve α_r by $\alpha_r(s) = \alpha(s, r)$. If α_r is timelike for all r, then we say α is a ***timelike variation***. Given a piecewise smooth variation α of a smooth timelike curve γ, we can always find a smaller variation of α such that α is a variation through timelike curves.

Proposition 7.5 *Let $\alpha : [a, b] \times (-\epsilon, \epsilon) \to M$ be a piecewise smooth variation of the piecewise smooth timelike curve $\gamma : [a, b] \to M$. Then there exists a $\delta > 0$ such that $\alpha|_{[a,b]\times(-\delta,\delta)}$ is a piecewise smooth timelike variation.*

Proof. First consider the case when α is smooth. Pick any $\epsilon_1 \in (0, \epsilon)$. Suppose there exists no $\delta > 0$ such that all the curves α_r are timelike for $|r| < \delta$. Then we can find a sequence $\{r_n\}$ converging to 0 such that the curves α_{r_n} failed to be timelike at some point s_n. This means $g\big(\alpha'_{r_n}(s_n), \alpha'_{r_n}(s_n)\big) \geq 0$. Since $[a, b] \times [-\epsilon_1, \epsilon_1]$ is compact, $\{(s_n, r_n)\}$ has a limit point (s_0, r_0) and since $r_n \to 0$, we must have $r_0 = 0$. Thus $g\big(\alpha'_0(s_0), \alpha'_0(s_0)\big) = g\big(\gamma'(s_0), \gamma'(s_0)\big) \geq 0$ which contradicts γ being timelike.

Now let α is a piecewise smooth variation of γ and $a = s_0 < s_1 < \cdots < s_k = b$ be the finite partition of $[a, b]$. For each i, $\alpha|_{[s_i,s_{i+1}]}$ is a smooth variation of $\gamma|_{[s_i,s_{i+1}]}$. From the above paragraph, for each i there exists a δ_i such that for $|r| < \delta_i$, $\alpha_r|_{[s_i,s_{i+1}]}$is timelike. Taking $\delta = \min\{\delta_i : 0 \leq i \leq k\}$ yields the required δ. □

Let $\alpha : [a, b] \times (-\epsilon, \epsilon) \to M$ be a variation of a timelike curve γ and let (s, r) be the standard coordinates for $[a, b] \times (-\epsilon, \epsilon)$. We define w to be the push forward of $\partial/\partial r$ under α and u to be the push forward of $\partial/\partial s$ under α. This means for each $s \in [s_{i-1}, s_i]$,

$w(s,r)$ and $u(s,r)$ are given by

$$w(s,r) = \big(\alpha|_{[s_i,s_i]\times(-\epsilon,\epsilon)}\big)_* \frac{\partial}{\partial r}\bigg|_{(s,r)}$$
$$u(s,r) = \big(\alpha|_{[s_i,s_i]\times(-\epsilon,\epsilon)}\big)_* \frac{\partial}{\partial s}\bigg|_{(s,r)}.$$

$w(s,0)$ is called the ***deviation vector*** of γ. We will sometimes abuse notation and write $u(s) = u(s,0)$ and $w(s) = w(s,0)$. Hence $w(s)$ is the deviation vector of γ. Given any piecewise smooth vector field w on γ, we can always find a piecewise smooth variation α of γ such that w is the deviation vector of γ.

Proposition 7.6 *Let $\gamma : [a,b] \to M$ be a smooth timelike curve and let w be any piecewise smooth vector field along γ. There exists a piecewise smooth variation $\alpha : [a,b]\times(-\epsilon,\epsilon) \to M$ of γ such that w is the deviation vector of γ. If $w(a) = w(b) = 0$, then the variation can be made proper.*

Proof. Let h be an auxiliary complete Riemannian metric on M. Consider a normal neighborhood $N(s)$ about each point $\gamma(s)$ for $s \in [a,b]$. Within each $N(s)$ we can find a ball $B(s) \subset N(s)$ centered around $\gamma(s)$ and with radius $\delta(s) > 0$ where the radius in terms of the Riemannian metric h. What this means is that the preimage of $B(s)$ under $\exp_{\gamma(s)}$ is an open ball in $T_{\gamma(s)}M$ contained in the preimage of $N(s)$ and this open ball has radius less than $\delta(s)$ where the radius is in terms of the Riemannian metric h. Since $\gamma\big([a,b]\big)$ is compact, we can find a finite cover $\{B(s_1),\ldots,B(s_n)\}$ for $\gamma\big([a,b]\big)$. Taking $\delta = \min_{1\le i\le n}\{\delta_i\}$, we see that $\exp_{\gamma(s)}$ is well-defined for all $s \in [a,b]$ and for all $v \in T_{\gamma(s)}M$ with $\sqrt{h(v,v)} < \delta$.

Define the number $N = \sup_{s\in[a,b]} \sqrt{h\big(w(s),w(s)\big)}$ (which is actually obtained for some $s \in [a,b]$ since the function is continuous) and fix $0 < \epsilon < \delta/N$. Then we define our variation by

$$\alpha(s,r) = \exp_{\gamma(s)}\big(rw(s)\big).$$

α is piecewise smooth since w is piecewise smooth. Moreover, the deviation vector (which is the push forward of $\partial/\partial r$ under α at $r = 0$) equals w since the differential of the exponential map is just the identity:

$$\alpha_*\frac{\partial}{\partial r}(s,0) = \frac{d}{dr}\big(\exp_{\gamma(s)} w(s)\big)\Big|_{r=0} = (d\exp_{\gamma(s)})_0 w(s) = w(s).$$

Hence w is the deviation vector and it follows that if $w(a) = w(b) = 0$, then α is proper. □

We will now derive the first and second variational formulas for the lorentzian distance functional. Set

$$\Delta_{s_i}u = \lim_{s\to s_i^+} u(s) - \lim_{s\to s_i^-} u(s) \quad \text{for } i = 1,2,\ldots,k-1,$$

$$\Delta_{s_k}u = -\lim_{s\to b^-} u(s), \quad \text{and} \quad \Delta_{s_0}(u) = \lim_{s\to a^+} u(s).$$

Recall from section 6.7 that if $\gamma : [a,b] \to M$ is a causal curve, then the length of γ is $L(\gamma) = \int_a^b \sqrt{-g\big(\gamma'(s),\gamma'(s)\big)ds}$ and $L(\gamma) = c\tau(\gamma)$ where $\tau(\gamma)$ is the proper time of γ. Let L be the length functional from sectioin 6.7 and put $L(r) = L(\alpha_r)$. Recall that $L(r)$ is just c times the proper time of α_r. We derive the first variational formula for L.

Proposition 7.7 *Let $\gamma : [a,b] \to M$ be a piecewise smooth timelike curve normalized such that $g(\gamma',\gamma') = -c^2$. Let $\alpha : [a,b]\times(-\epsilon,\epsilon) \to M$ be a timelike variation of γ with w*

and u being the push forward of $\partial/\partial r$ and $\partial/\partial s$, respectively. Then

$$\begin{aligned}\left.\frac{dL}{dr}\right|_{r=0} &= \frac{1}{c}\int_a^b g\left(w, \nabla_u u\right)\big|_{r=0} ds + \frac{1}{c}\sum_{i=0}^{k} g\big(w(s_i), \Delta_{s_i}(u)\big) \\ &= \frac{1}{c}\int_a^b \left[w_a u^b \nabla_b u^a\right]_{r=0} ds + \frac{1}{c}\sum_{i=0}^{k} w^a(s_i)(\Delta_{s_i} u)_a.\end{aligned}$$

Proof. Let $L_i : (-\epsilon, \epsilon) \to \mathbb{R}$ denote the arc length function of $\alpha_r|_{[s_{i-1}, s_i]}$, then $L(r) = \sum_{i=1}^{k} L_i(r)$. Let us focus on each L_i. We have

$$\begin{aligned}\frac{dL_i}{dr} &= \frac{d}{dr}\int_{s_{i-1}}^{s_i} \sqrt{-u^a u_a}\, ds \\ &= \int_{s_{i-1}}^{s_i} \frac{1}{2\sqrt{-u^c u_c}}(-2u_a w^b \nabla_b u^a) \\ &= \int_{s_{i-1}}^{s_i} \frac{-1}{\sqrt{-u^c u_c}} u_a u^b \nabla_b w^a.\end{aligned}$$

The second equality follows from the chain rule and the third equality follows from $[u, w] = u^b \nabla_b w^a - w^b \nabla_b u^a = 0$ since u and v are the push forward of coordinate vector fields. Now at $r = 0$, we have $u^a u_a = -c^2$. Also, $\frac{d}{ds}(u^a w_a) = u_a u^b \nabla_b w^a + w_a u^b \nabla_b u^a$. Therefore

$$\begin{aligned}\left.\frac{dL_i}{dr}\right|_{r=0} &= \frac{1}{c}\int_{s_{i-1}}^{s_i} \left[w_a u^b \nabla_b u^a\right]_{r=0} ds - \frac{1}{c}\int_{s_{i-1}}^{s_i} \frac{d}{ds}[u^a w_a]_{r=0} ds \\ &= \frac{1}{c}\int_{s_{i-1}}^{s_i} \left[w_a u^b \nabla_b u^a\right]_{r=0} ds - \frac{1}{c}[u^a w_a]_{r=0}\Big|_{s_{i-1}^-}^{s_i^+}.\end{aligned}$$

In order for α to be continuous, we must have $w_a(s_i^+) = w_a(s_i^-)$, so summing over all i, we have

$$\left.\frac{dL}{dr}\right|_{r=0} = \frac{1}{c}\int_a^b [w_a u^b \nabla_b u^a]_{r=0} ds + \frac{1}{c}\sum_{i=0}^{k} [w^a(s_i)(\Delta_{s_i} u)_a]_{r=0},$$

which is the desired formula. □

Corollary 7.8 *Let $\alpha : [a,b] \times (-\epsilon, \epsilon)$ be a variation of the timelike geodesic γ (i.e. $\alpha_0(s) = \gamma(s)$). Then*

$$\left.\frac{dL}{dr}\right|_{r=0} = \frac{1}{c} g(\alpha_0', \gamma')\Big|_a^b.$$

Proof. Since γ is a geodesic, $\nabla_{\gamma'}\gamma' = 0$ and it's at least C^2. Therefore $\Delta_{s_i}\gamma' = \lim_{s\to s_i^+} \gamma'(s) - \lim_{s\to s_i^-} \gamma'(s) = 0$ except for $i = 0$ and $i = k$. □

Now we derive the more complicated second variational formula for $L(r)$. For w and u defined as above, we define the vector field n as $v^a = c^2 w^a + w^b u_b u^a$. Hence if w and u are orthogonal everywhere, then $v = c^2 w$. Also, if u is normalized such that $u^a u_a = -c^2$, then $v^a u_a = 0$. Like u, put $v(s) = v(s,0)$ and define

$$\Delta_{s_i} v' = \lim_{s\to s_i^+} \nabla_u v(s) - \lim_{s\to s_i^-} \nabla_u v(s) \quad \text{for } i = 1, 2, \ldots, k-1,$$

$$\Delta_{s_k} v' = -\lim_{s\to b^-} \nabla_u v(s), \quad \text{and} \quad \Delta_{s_0}(v') = \lim_{s\to a^+} \nabla_u v(s).$$

Proposition 7.9 *Let $\gamma : [a,b] \to M$ be a smooth (at least C^2) geodesic normalized such that $g(\gamma', \gamma') = -c^2$. Let $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$ be a timelike variation of γ with w and u being the push forward of $\partial/\partial r$ and $\partial/\partial s$, respectively, and define the vector field v as $v^a = c^2 w^a + w^b u_b u^a$. Then*

$$\begin{aligned}
\left.\frac{d^2 L}{dr^2}\right|_{r=0} &= \frac{1}{c^5}\int_a^b g\big(v, \nabla_u(\nabla_u v) - c^2 R(u,w)u\big)\big|_{r=0} ds \\
&\quad + \frac{1}{c^5}\sum_{i=0}^{k} g\big(v(s_i), \Delta_{s_i} v'\big) - \frac{1}{c} g(u, \nabla_w w)_{r=0}\Big|_a^b \\
&= \frac{1}{c^5}\int_a^b \big[v_a\big(u^b \nabla_b (u^d \nabla_d v^a) - c^2 R^a{}_{cbd} w^d u^b u^c\big)\big]_{r=0} ds \\
&\quad + \frac{1}{c^5}\sum_{i=0}^{k} v^a(s_i)(\Delta_{s_i} v')_a - \frac{1}{c}[u_a w^b \nabla_b w^a]_{r=0}\Big|_a^b.
\end{aligned}$$

Proof. Just like in the proof of proposition 7.7, we restrict attention to $L_i(r)$ of $\alpha_r|_{[s_{i-1},s_i]}$. Recall that

$$\frac{dL_i}{dr} = \int_{s_{i-1}}^{s_i} \frac{-1}{\sqrt{-u^c u_c}}(u_a u^b \nabla_b w^a) ds,$$

so differentiating once more, we get

$$\begin{aligned}\frac{d^2 L_i}{dr^2} &= \int_{s_{i-1}}^{s_i} \frac{d}{dr}\left[\frac{-1}{\sqrt{-u^c u_c}}(u_a u^b \nabla_b w^a)\right] ds \\ &= \int_{s_{i-1}}^{s_i} \left[-\frac{(u_a u^b \nabla_b u^a)^2}{(-u^c u_c)^{3/2}} - \frac{(w^d \nabla_d u_a)(u^b \nabla_b w^a) + u_a w^d \nabla_d (u^b \nabla_b w^a)}{(-u^c u_c)^{1/2}}\right] ds. \qquad (7.11)\end{aligned}$$

Now we will start calculating relevant quantities at $r = 0$. Recall we put $u(s) = u(s, 0) = \gamma'(s)$. Then since γ is a geodeisc, we have at $r = 0$ and $s \in (s_{i-1}, s_i)$

$$u^b \nabla_b (w^c u_c u^a) = u^a u^b \nabla_b (w^c u_c) + w^c u_c u^b \nabla_b u^a = u^a u^b \nabla_b (w^c u_c).$$

Also, at $r = 0$ and $s \in (s_{i-1}, s_i)$ we have $v^a u_a = 0$, so

$$u_a u^b \nabla_b v^a = u^b \nabla_b (v_a u^a) - v_a u^b \nabla_b u^a = 0.$$

These last two calculations imply

$$(u^b \nabla_b v_a) u^d \nabla_d (w^c u_c u^a) = (u^b \nabla_b v_a) u^a u^b \nabla_b (w^c u_c) = 0.$$

Now we focus on the term $(w^d\nabla_d u_a)(u^b\nabla_b w^a)$ in eq. (7.11). Since $c^2 w = v - g(w,u)u$, we have at $r = 0$ and $s \in (s_{i-1}, s_i)$

$$\begin{aligned}
c^4(w^d\nabla_d u_a)(u^b\nabla_b w^a) &= c^4(u^d\nabla_d w_a)(u^b\nabla_b w^a) \\
&= u^b\nabla_b(v^a - w^c u_c u^a)u^d\nabla_d(v_a - w^c u_c u_a) \\
&= (u^b\nabla_b v^a)(u^d\nabla_d v_a) - 2(u^b\nabla_b v^a)u^d\nabla_d(w^c u_c u_a) \\
&\quad + u^b\nabla_b(w^c u_c u^a)u^d\nabla_d(w^c u_c u_a) \\
&= (u^b\nabla_b v^a)(u^d\nabla_d v_a) + u^b\nabla_b(w^c u_c u^a)u^d\nabla_d(w^c u_c u_a) \\
&= (u^b\nabla_b v^a)(u^d\nabla_d v_a) + \big(u^a u^b\nabla_b(w^c u_c)\big)\big(u_a u^b\nabla_b(w^c u_c)\big) \\
&= (u^b\nabla_b v^a)(u^d\nabla_d v_a) - c^2\big(u^b\nabla_b(w^c u_c)\big)^2 \\
&= (u^b\nabla_b v^a)(u^d\nabla_d v_a) - c^2\big(u_c u^b\nabla_b w^c\big)^2.
\end{aligned}$$

Calculating eq. (7.11) at $r = 0$ and using the above calculation, we find

$$\left.\frac{d^2 L_i}{dr^2}\right|_{r=0} = \int_{s_{i-1}}^{s_i} \left[-c^{-5}(u^b\nabla_b v^a)(v^d\nabla_d v_a) - c^{-1}u_a w^d\nabla_d(u^b\nabla_b w^a)\right] ds \tag{7.12}$$

Now we focus on $w^d\nabla_d(u^b\nabla_b w^a) = \big[\nabla_w(\nabla_u w)\big]^a$. Since $[u,w] = 0$, the Riemann tensor gives

$$\nabla_w(\nabla_u w) = R(w,u)w + \nabla_u(\nabla_w w).$$

In components $w^d\nabla_d(u^b\nabla_b w^a) = R^a{}_{cdb}w^d u^b w^c + u^d\nabla_d(w^b\nabla_b w^a)$. Plugging this into eq. (7.12), we find

$$\begin{aligned}
\left.\frac{d^2 L_i}{dr^2}\right|_{r=0} &= \int_{s_{i-1}}^{s_i} \left[-c^{-5}(u^b\nabla_b v^a)(v^d\nabla_d v_a) - c^{-1}u_a\big(R^a{}_{cdb}w^d u^b w^c + u^d\nabla_d(w^b\nabla_b w^a)\big)\right] ds \\
&= \int_{s_{i-1}}^{s_i} \left[-c^{-5}(u^b\nabla_b v^a)(v^d\nabla_d v_a) - c^{-1}u_a R^a{}_{cdb}w^d u^b w^c - c^{-1}u^d\nabla_d(u_a w^b\nabla_b w^a)\right] ds.
\end{aligned}$$

The last equality follows since γ is a geodesic. Now $u^d\nabla_d(u_a w^b \nabla_b w^a) = \frac{d}{ds} g(u, \nabla_w w)$. Therefore we can integrate that term out using the fundamental theorem of calculus. We find

$$\left.\frac{d^2 L_i}{dr^2}\right|_{r=0} = \int_{s_{i-1}}^{s_i} \left[-c^{-5}(u^b\nabla_b v^a)(v^d \nabla_d v_a) - c^{-1} u_a R^a{}_{cdb} w^d u^b w^c\right] ds - c^{-1}(u_a w^b \nabla_b w^a)\Big|_{s_{i-1}}^{s_i}. \tag{7.13}$$

Now let's work on $(u^b\nabla_b v^a)(u^d\nabla_d v_a) = g(\nabla_u v, \nabla_u v)$. We find

$$(u^b\nabla_b v^a)(u^d\nabla_d v_a) = u^b\nabla_b(v^a u^d \nabla_d v_a) - v^a u^b \nabla_b u^d \nabla_d v_a.$$

Like before, we find $u^b\nabla_b(v^a u^d\nabla_d v_a) = \frac{d}{ds} g(v, \nabla_u v)$. So applying the fundamental theorem of calculus again, we find

$$\begin{aligned}\left.\frac{d^2 L_i}{dr^2}\right|_{r=0} &= \int_{s_{i-1}}^{s_i} \left[c^{-5} v_a u^b \nabla_b u^d \nabla_d v^a - c^{-1} u_a R^a{}_{cdb} w^d u^b w^c\right] ds \\ &\quad - c^{-5} v_a u^d \nabla_d v^a\Big|_{s_{i-1}}^{s_i} - c^{-1}(u_a w^b \nabla_b w^a)\Big|_{s_{i-1}}^{s_i}.\end{aligned}$$

Thus summing over all $i = 1, 2, \ldots, k$ and recalling that γ is a smooth geodesic, we find

$$\begin{aligned}\left.\frac{d^2 L}{dr^2}\right|_{r=0} &= \sum_{i=1}^{k} \frac{d^2 L_i}{dr^2} \\ &= \int_a^b \left[c^{-5} v_a u^b \nabla_b u^d \nabla_d v^a - c^{-1} u_a R^a{}_{cdb} w^d u^b w^c\right]_{r=0} ds \\ &\quad + c^{-5} \sum_{i=0}^{k} [v^a(s_i)(\nabla_{s_i} v')_a]_{r=0} - c^{-1}[u_a w^b \nabla_b w^a]_{r=0}\Big|_a^b.\end{aligned} \tag{7.14}$$

Now we use the curvature property $R_{abcd} = R_{badc}$ to get

$$\begin{aligned} u_a R^a{}_{cdb} w^d u^b w^c &= R_{acdb} w^d u^b w^c u^a \\ &= R_{cabd} w^d u^b w^c u^a \\ &= w_c R^c{}_{abd} w^d u^b u^a \\ &= c^{-2}(v_c - w^e u_e u_c) R^c{}_{abd} w^d u^b u^a \\ &= c^{-2} v_c R^c{}_{abd} w^d u^b u^a. \end{aligned}$$

In the last equality we used $u_c R^c{}_{abd} w^d u^b u^a = R_{bdca} u^c w^d u^b u^a = -R_{dbca} u^c w^d u^b u^a = 0$ since $R(u,u)u = 0$. Substituting this into eq. (7.14), we find

$$\begin{aligned} \left.\frac{d^2 L}{dr^2}\right|_{r=0} = \int_a^b c^{-3} \left[v_a (c^{-2} u^b \nabla_b u^d \nabla_d v^a - R^a{}_{cbd} w^d u^b u^c) \right]_{r=0} ds \\ + c^{-5} \sum_{i=0}^{k} [v^a(s_i)(\Delta_{s_i} v')_a]_{r=0} - c^{-1} [u_a w^b \nabla_b w^a]_{r=0} \Big|_a^b, \end{aligned}$$

which is what we wanted to show. □

We derived the second variational formula for any arbitrary deviation vector $w(s) = w(s,0)$. We will now restrict our attention to deviation vectors w which are orthogonal to γ. For a timelike curve γ, we define $V^\perp(\gamma)$ to be the infinite-dimensional vector space of all piecewise smooth vector fields w along γ such that $g(w, \gamma') = 0$ everywhere along γ. Recall that we defined v by $v = c^2 w + g(w,u)u$. So for $w\big|_{r=0} \in V^\perp(\gamma)$, we have $v = c^2 w$ along γ. In this case, proposition 7.9 implies

Corollary 7.10 *Let $\gamma : [a,b] \to M$ be a smooth (at least C^2) geodesic normalized such that $g(\gamma', \gamma') = -c^2$. Let $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$ be a timelike variation of γ with w and*

u being the push forward of $\partial/\partial r$ and $\partial/\partial s$, respectively, with $w\big|_{r=0} \in V^{\perp}(\gamma)$. Then

$$\left.\frac{d^2L}{dr^2}\right|_{r=0} = \frac{1}{c}\int_a^b g\big(w, \nabla_u(\nabla_u w) - R(u,w)u\big)\big|_{r=0} ds + \frac{1}{c}\sum_{i=0}^{k} g\big(w(s_i), \Delta_{s_i} w'\big) - \frac{1}{c} g(u, \nabla_w w)_{r=0}\Big|_a^b.$$

Motivated by corollary (7.10), we define the ***Lorentzian Index Form*** as the symmetric bilinear form $I : V^{\perp}(\gamma) \times V^{\perp}(\gamma) \to \mathbb{R}$ given by

$$I(v,w) = -\frac{1}{c}\int_a^b \big[g(\nabla_{\gamma'} v, \nabla_{\gamma'} w) + g\big(R(\gamma', v)\gamma', w\big)\big] ds$$

for a timelike geodesic γ. The integral above is well-defined since the points of discontinuity of $\nabla_{\gamma'} v$ and $\nabla_{\gamma'} w$ is finite and thus a set of measure zero. Let $v \in V^{\perp}(\gamma)$ so that there is a partition $a = s_0 < s_1, \cdots < s_k = b$ and $v\big|_{[}s_i, s_{i+1}]$ is smooth for each $i = 0, 1, \ldots, k-1$. Then using the compatibility of the metric, we find

$$I(v,w) = \frac{1}{c}\int_a^b g\big(w, \nabla_{\gamma'}(\nabla_{\gamma'} v) - R(\gamma', v)\gamma'\big) ds + \frac{1}{c}\sum_{i=0}^{k} g\big(\Delta_{s_i} v', w(s_i)\big).$$

Notice that when set equal to zero, the integrand is precisely the geodesic deviation equation (Jacobi equation). By corollary (7.10) we see that the Lorentzian index form applied to the (w, w) is precisely the second variational formula for proper variations (i.e. $\alpha_r(a) = \gamma(a)$ and $\alpha_r(b) = \gamma(b)$ for all $r \in (-\epsilon, \epsilon)$).

Proposition 7.11 *Let $\gamma : [a,b] \to M$ be a smooth (at least C^2) geodesic normalized such that $g(\gamma', \gamma') = -c^2$. Let $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$ be a proper timelike variation of γ with w and u being the push forward of $\partial/\partial r$ and $\partial/\partial s$, respectivley, with $w\big|_{r=0} \in V^{\perp}(\gamma)$. Then*

$$\left.\frac{d^2L}{dr^2}\right|_{r=0} = I(w,w).$$

Proof. By corollary 7.10 it suffices to show $\frac{1}{c}g\big(\gamma', \nabla_w w(0)\big)\big|_a^b = 0$. This follows since, because α is proper, $w(0) = 0$ at a and b. $\square$

Therefore if we're given a proper timelike variation α with deviation vector w such that $I(w, w) > 0$, then we can find timelike curves α_r joining $\gamma(a)$ to $\gamma(b)$ with $L(r) = L(\alpha_r) > L(\gamma)$.

We are almost ready to prove that conjugate points mark the end of timelike geodesics maximizing proper time. But first we need the following lemma.

Lemma 7.12 *Let $\gamma : [a, b] \to M$ be a timelike geodesic and let w be any Jacobi field along γ. Then*

(a) *$g(w, \gamma')$ is an affine function of s, i.e. $g\big(w(s), \gamma'(s)\big) = \alpha s + \beta$ for some constants $\alpha, \beta \in \mathbb{R}$.*

(b) *if $w(s_1) = w(s_2) = 0$ for distinct $s_1, s_2 \in [a, b]$, then $w \in V^{\perp}(\gamma)$.*

(c) *if w is a Jacobi field which vanishes at a and b, then $\nabla_{\gamma'} w \in V^{\perp}(\gamma)$.*

Proof. We have to show $\frac{d^2}{ds^2} g(w, \gamma') = 0$. Using the compatibility of the metric and the fact that γ is a geodesic, we have

$$\frac{d^2}{ds^2} g(w, \gamma') = g\big(\nabla_{\gamma'}(\nabla_{\gamma'} w), \gamma'\big) = g\big(R(\gamma', w)\gamma', \gamma'\big) = -g\big(R(\gamma', \gamma')\gamma', w\big) = 0$$

This establishes (a). For (b) we see that if $w(s_1) = w(s_2) = 0$ for distinct s_1 and s_2, then $\alpha s_1 + \beta = \alpha s_2 + \beta = 0$. This holds only if $\alpha = \beta = 0$. (c) follows from (b) and the compatibility of the metric. $\square$

We are now ready to prove that conjugate points mark the end of timelike geodesics maximizing proper time. More precisely, we have

Theorem 7.13 *Suppose that $\gamma : [a, b] \to M$ is a timelike geodesic and some point $r = \gamma(s_0)$, $s_0 \neq a, b$, is conjugate to the point $\gamma(a)$. Then there exists a piecewise smooth proper variation $\alpha : [a, b] \times (-\epsilon, \epsilon) \to M$ of γ such that $L(\alpha_r) > L(\gamma)$ for all $r \neq 0$. Thus $\gamma : [a, b] \to M$ is not maximal.*

Proof. We seek a proper timelike variation α of γ such that $d^2L/dr^2\big|_{r=0} > 0$. Because, since α is proper, corollary 7.8 implies $dL/dr\big|_{r=0} = 0$ so γ is a critical point of the length functional and so $d^2L/dr^2\big|_{r=0}$ implies γ is a local minimum so the timelike curves α_r for r close to 0 will all have lengths longer than γ. By proposition 7.11, it suffices to find a timelike proper variation with deviation vector w orthogonal γ and satisfying $I(w, w) > 0$. Given $w \in V^{\perp}(\gamma)$ which vanishes at a and b, we can construct the desired timelike variation α by propositions 7.6 and 7.5. Thus it suffices to find a vector field $w \in V^{\perp}(\gamma)$ such that $w(a) = w(b) = 0$ and $I(w, w) > 0$.

Since r is conjugate to $\gamma(a)$, there exists a nontrivial Jacobi field w_1 along γ which vanishes at $\gamma(a)$ and r. By lemma 7.12, we have $w_1 \in V^{\perp}(\gamma)$ and $\nabla_{\gamma'} w_1 \in V^{\perp}(\gamma)$. Since $w_1(s_0) = 0$ but w_1 is nontrivial, $\nabla_{\gamma'} w_1(s_0)$ is a *nonzero* spacelike vector.

We will denote $I(\cdot, \cdot)_a^{s_0}$ the restriction of the Lorentzian index form to $\gamma\big|_{a,s_0}$. Then

for any $v \in V^{\perp}(\gamma)$, we have

$$\begin{aligned} I(w_1, v)_a^{s_0} &= -\frac{1}{c}\int_a^{s_0} \big[g\big(\nabla_{\gamma'} w_1, \nabla_{\gamma'} v\big) + g\big(R(\gamma', w_1)\gamma', v\big)\big] ds \\ &= -\frac{1}{c} g\big(\nabla_{\gamma'} w_1, v\big)\big|_a^{s_0} + \frac{1}{c}\int_a^{s_0} g\big(v, \nabla_{\gamma'}(\nabla_{\gamma'} w_1) - R(\gamma', w_1)\gamma'\big) ds \\ &= -\frac{1}{c} g\big(\nabla_{\gamma'} w_1, v\big)\big|_a^{s_0}. \end{aligned} \tag{7.15}$$

The second equality follows from using the compatibility of the metric and the fact that w_1 is at least C^2 being a Jacobi field. The third equality follows since $\nabla_{\gamma'}(\nabla_{\gamma'} w_1) - R(\gamma', w_1)\gamma' = 0$ since w_1 is a Jacob field.

We will now construct a piecewise smooth vector field $w \in V^{\perp}(\gamma)$ such that $w(a) = w(b) = 0$ and $I(w, w) > 0$. Let $\psi : [a, b] \to \mathbb{R}$ be a smooth function with $\psi(a) = \psi(b) = 0$ and $\psi(s_0) = 1$. Let v_1 be a smooth parallel vector field along γ with $v_1(s_0) = -\nabla_{\gamma'} w_1(s_0)$ which, recall, is nonzero. Then put $v = \psi v_1$. Recognize that $v \in V^{\perp}(\gamma)$ and $v(a) = v(b) = 0$ by ψ. For $\epsilon > 0$ define the one-parameter family w_ϵ by

$$w_\epsilon(s) = \begin{cases} w_1(s) + \epsilon v(s) & : \text{ for } a \le s \le s_0 \\ \epsilon v(s) & : \text{ for } s_0 < s \le b. \end{cases}$$

Using eq. (7.15), we have

$$\begin{aligned} I(w_\epsilon, w_\epsilon) &= I(w_\epsilon, w_\epsilon)_a^{s_0} + I(w_\epsilon, w_\epsilon)_{s_0}^{b} \\ &= I(w_1 + \epsilon v, w_1 + \epsilon v)_a^{s_0} + I(\epsilon v, \epsilon v)_{s_0}^{b} \\ &= I(w_1, w_1)_a^{s_0} + 2\epsilon I(w_1, v)_a^{s_0} + \epsilon^2 I(v, v)_a^{s_0} + \epsilon^2 I(v, v)_{s_0}^{b} \\ &= -\frac{1}{c} g\big(\nabla_{\gamma'} w_1, w_1\big)\big|_a^{s_0} - \frac{2\epsilon}{c} g\big(\nabla_{\gamma'} w_1, v\big)\big|_a^{s_0} + \epsilon^2 I(v, v) \\ &= -\frac{2\epsilon}{c} g\big(\nabla_{\gamma'} w_1(s_0), v(s_0)\big) + \epsilon^2 I(v, v) \\ &= \frac{2\epsilon}{c} g\big(\nabla_{\gamma'} w_1(s_0), \nabla_{\gamma'} w_1(s_0)\big) + \epsilon^2 I(v, v). \end{aligned}$$

The third equality uses the linearity of the Lorentzian index form, the fourth equality uses eq. (7.15), the fifth equality uses $w_1(a) = w_1(s_0) = v(a) = 0$, and the sixth equality uses the fact that $v(s_0) = v_1(s_0) = -\nabla_{\gamma'} w_1(s_0)$. Recall that $\nabla_{\gamma'} w_1(s_0)$ is nonzero and spacelike, so $\frac{2\epsilon}{c} g\big(\nabla_{\gamma'} w_1(s_0), \nabla_{\gamma'} w_1(s_0)\big) > 0$, so if $I(v,v) > 0$, then $I(w_\epsilon, w_\epsilon) < 0$ for any $\epsilon > 0$. If $I(v,v)$ is negative, then pick $\epsilon > 0$ such that

$$0 < \epsilon < -\frac{2g\big(\nabla_{\gamma'} w_1(s_0), \nabla_{\gamma'} w_1(s_0)\big)}{cI(v,v)}.$$

Then $I(w_\epsilon, w_\epsilon) > 0$. □

7.4 Timelike Variations Applied to Focal Points

In section 7.2 we defined the notion of what it meant for a point q to be conjugate to another point p, and in section 7.3 we used variational principles to show that conjugate points mark the end of timelike geodesics maximizing proper time which was stated more precisely in Theorem 7.13. In this section, we will define what it means for a point q to be conjugate to a spacelike hypersurface Σ. However, before we do this let us review some background of the second fundamental form.

Let N be a smooth (at least C^2) submanifold of M without boundary. We will only consider submanifolds that are boundaryless. Let $i : N \to M$ be the inclusion map and identify $di_p(T_pN)$ with T_pN so that we can regard T_pN as being a subspace of T_pM. Let $g_0 = i^*g$ be the pullback of the Lorentzian metric g on M to a symmetric tensor field g_0 at p. Using the identification of T_pN with $di_p(T_pN)$, we also identify g_0 at p with $g\big|_{T_pN \times T_pN}$ at p.

We will assume N is ***nondegenerate***, that is, for each $p \in N$ and nonzero $v \in T_pN$, there exists a $w \in T_pN$ such that $g(v,w) \neq 0$. If $g\big|_{T_pN \times T_pN}$ is positive definite for all $p \in N$, then N is said to be ***spacelike***. If $g\big|_{T_pN \times T_pN}$ is a Lorentzian metric for each $p \in N$, then N is ***timelike***. We define

$$T_p^{\perp}N = \{v \in T_pM : g(v,w) = 0 \text{ for all } w \in T_pN\}.$$

Since we're assuming N is nondegenerate, we have $T_p^{\perp}N \cap T_pN = \{0\}$. Therefore $T_pM = T_p^{\perp}N \oplus T_pN$. So given any vector $v \in T_pM$, we can uniquely decompose $v = v^{\perp} + v^N$ such that $v^{\perp} \in T_p^{\perp}N$ and $v^N \in T_pN$. This allows us to define the orthogonal projection map $P : T_pM \to T_pN$ by $P(v) = v^N$. Let ∇ be the unique torsion free derivative operator which is compatible with g. We define the connection $\nabla^N = P \circ \nabla$ for tensor fields defined on N (e.g. $\nabla^N_v w = P(\nabla_v w)$ for $v, w \in TN$). ∇^N is torsion-free follows from ∇ being torsion free. Moreover ∇^N is compatible with g_0 as the following calculation shows. Pick u, v, and w smooth in TN.

$$\begin{aligned} u\big(g_0(v,w)\big) &= u\big(g(v,w)\big) \\ &= g(\nabla_u v, w) + g(v, \nabla_u w) \\ &= g_0(\nabla^N_u v, w) + g_0(v, \nabla^N_u w) \end{aligned}$$

where the last equality follows since v and w are in TN. Thus ∇^N is the unique torsion free derivative operator which is compatible with g_0.

Given $n \in T^{\perp}N = \bigcup_{p \in N} T^{\perp}N$, we define the ***second fundamental form*** as the map $S_n : TN \times TN \to C^{\infty}(N)$ in the following way. Extend the vectors $v, w \in TN$ to

vector fields $\widetilde{v}, \widetilde{w} \in TM$ such that $v = \widetilde{v}$ and $w = \widetilde{w}$ on N. Then

$$S_n(v,w) = g\big(\nabla_{\widetilde{v}}\widetilde{w}, n\big) = g\big((\nabla_{\widetilde{v}}\widetilde{w})^{\perp}, n\big) = \big(\nabla_{\widetilde{v}}\widetilde{w} - \nabla^N_{\widetilde{v}}\widetilde{w}, n\big).$$

Proposition 7.14 *Given $n \in T^{\perp}_p N$, S_n is symmetric, bilinear, and does not depend on the extensions used for its arguments.*

Proof. We first show S_n is symmetric. We have

$$\begin{aligned} S_n(v,w) - S_n(w,v) &= g\big(\nabla_{\widetilde{v}}\widetilde{w} - \nabla_{\widetilde{w}}\widetilde{v} - \nabla^N_{\widetilde{v}}\widetilde{w} + \nabla^N_{\widetilde{w}}\widetilde{v}, n\big) \\ &= g\big(\nabla_{\widetilde{v}}\widetilde{w} - \nabla_{\widetilde{w}}\widetilde{v}, n\big) \\ &= g\big([\widetilde{v}, \widetilde{w}], n\big). \end{aligned}$$

It suffices to show $[\widetilde{v}, \widetilde{w}]$ is tangent to N. This follows immediately from the following fact: $\widetilde{v}$ is tangent to M if and only if $\widetilde{v}(f) = 0$ for all smooth functions f which vanish on N. To prove this fact, choose a basis for TM consisting of a basis for $T^{\perp}N$ union a basis for TN and calculate $\widetilde{v}(f)$ in components with respect to this basis. This show S_n is symmetric.

To show S_n doesn't depend on the extensions, notice that S_n does not depend on the extension given for v by the tensoral properties of ∇. So using the symmetry just proved, S_n also doesn't depend on the extension $\widetilde{w}$. This argument also shows S_n is bilinear. $\square$

Just like in the Riemannian case, one can show that a submanifold is totally geodesic if and only if the second fundamental form vanishes on N for all $n \in T^{\perp}N$. However, it is not needed for our purposes, so we omit it.

Since S_n doesn't depend on the extensions used, we will abuse notation and write $S_n(v,w) = g(\nabla_v w, n)$. Given $n \in T^\perp N$, we define the ***second fundamental form operator*** $L_n : TN \to TN$ by $g\big(L_n(v), w\big) = S_n(v,w) = g(\nabla_v w, n)$ for all $v, w \in TN$. For a spacelike hypersurface Σ, we can characterize L_n in the following way.

Lemma 7.15 *Let Σ be a spacelike hypersurface with timelike normal field n such that $g(n,n) = -c^2$. If v is a tangent vector for Σ, then $L_n(v) = -\nabla_v n$. Hence the components of L_n are given by $(L_n)^a{}_b = -\nabla_b n^a$.*

Proof. Since $g(n,n) = -c^2$ is constant, we have $0 = v\big(g(n,n)\big) = 2g\big(\nabla_v n, n\big)$ which shows that $\nabla_v n \in T\Sigma$. If w is any vector field tangent to Σ, then $g(n,w) = 0$. Therefore

$$0 = v\big(g(n,w)\big) = g(\nabla_v n, w) + g(n, \nabla_v w)$$

which shows $g(n, \nabla_v w) = g(-\nabla_v n, w)$. Hence

$$g\big(L_n(v), w\big) = g\big(\nabla_v w, n\big) = g(-\nabla_v n, w).$$

Since this is true for all w, we have $L_n(v) = -\nabla_v n$. □

Let us now consider a spacelike hypersurface Σ with $n \in T^\perp\Sigma$ normalized such that $g(n,n) = -c^2$. The collection of timelike geodesics orthogonal to Σ with initial direction n forms a congruence of timelike geodesics. Let γ be a timelike geodesic in this congruence which intersects Σ at p and let w be a vector field along γ which solves the Jacobi equation. By the previous Lemma, we see that w satisfies the following condition on Σ

$$\nabla_{\gamma'} w = \nabla_w \gamma' = \nabla_w n = -L_n(w).$$

This motivates the following definition. Let γ be a timelike geodesic which is orthogonal to the spacelike hypersurface Σ at p. A point q on γ is said to be a ***focal point of*** Σ ***along*** γ if there is a nontrivial Jacobi field w along γ such that w is orthogonal to γ, vanishes at q, and satisfies $\nabla_{\gamma'} w = -L_n(w)$ at p. The same proof used in proposition 7.3 establishes the following analogous result.

Proposition 7.16 *Consider the timelike geodesic congruence which emanates orthogonally from a spacelike hypersurface Σ. Let γ be one of the timelike geodesics. Then $q \in \gamma$ is a focal point of Σ along γ if and only if the expansion θ of the congruence approaches $-\infty$ along γ.*

Proof. Mimic the proof of proposition 7.3.

This combined with lemma 7.2 yields the analogous result of proposition 7.4.

Proposition 7.17 *Let u describe the timelike geodesic congruence which emanates orthogonally from a spacelike hypersurface Σ with unit normal n and suppose $R_{ab}u^a u^b \geq 0$ (which will be the case if the strong energy condition is satisfied). If $-\mathrm{tr}(L_n) = -(L_n)^a{}_a = \nabla_a n^a = \theta_0 < 0$ is negative at some point $p \in \Sigma$, then within proper time $\tau \leq 3/|\theta_0|$ there is a focal point q to Σ along the geodesic γ orthogonal to Σ which starts at p, provided γ can be extended that far.*

Proof. First recall that the congruence u is irrotational by lemma 7.1. Along Σ, the expansion of the congruence is $\theta = \nabla_a u^a = \nabla_a n^a = -\mathrm{tr}(L_n)$. Since $-\mathrm{tr}(L_n) = \theta_0$ at p,

by lemma 7.2, $\theta \to -\infty$ along the geodesic γ in the congruence which starts at p. By proposition 7.16 there is a point $q \in \gamma$ which is a focal point of Σ, provided γ can be extended that far. $\square$

Now we want to show that focal points mark the end of timelike geodesics from Σ maximizing poper time, that is, we want to formulate an analogous theorem to that of theorem 7.13. To do this, we need to understand the first and second variational formulas of the length functional with regards to variations which start on spacelike hypersurfaces and end at a common point. More precisely, given a spacelike hypersurface Σ, we consider variations $\alpha : [a, b] \times (-\epsilon, \epsilon) \to M$ of a timelike curve $\gamma : [a, b] \to M$ such that $\alpha(a, r) \in \Sigma$ and $\alpha(b, r) = \gamma(b)$ for all $r \in (-\epsilon, \epsilon)$. The following proposition shows that, for at least our purposes, we need only consider variations of timelike geodesics.

Proposition 7.18 *Let Σ be a spacelike hypersurface. If $\gamma : [a, b] \to M$ is a piecewise timelike curve from Σ to a point $q = \gamma(b) \in M$, then a necessary and sufficient condition for γ to have maximal length among all timelike curves from Σ to q is that γ is a timelike geodesic which is orthogonal to Σ at $p = \gamma(a)$.*

Proof. Let $\alpha : [a, b] \times (-\epsilon, \epsilon) \to M$ be a timelike variation of γ such that $\alpha(a, r) \in \Sigma$ and $\alpha(b, r) = \gamma(b)$ for all $r \in (-\epsilon, \epsilon)$ and let $L(r) = L(\alpha_r)$ denote the length functional. Since γ has maximal length, we have $dL/dr\big|_{r=0} = 0$. Letting w and u be the push forward of $\partial/\partial r$ and $\partial/\partial s$, respectively, we have from proposition 7.7

$$0 = \left.\frac{dL}{dr}\right|_{r=0} = \frac{1}{c}\int_a^b g\left(w, \nabla_u u\right)\big|_{r=0} ds + \frac{1}{c}\sum_{i=1}^{k-1} g\big(w(s_i), \Delta_{s_i}(u)\big) + \frac{1}{c} g\big(w(a), u(a)\big).$$

The sum runs only to $k-1$ instead of k because $w(b) = w(b,0) = 0$. For $i = 1, \ldots, k-1$,

$$\Delta_{s_i}(u) = \lim_{s \to s_i^+} u(s) - \lim_{s \to s_i^-} u(s).$$

Assuming γ is maximal, each $\Delta_{s_i} = 0$ for if there existed an i such that $\Delta_{s_i} \neq 0$, then using a normal neighborhood about $\gamma(s_i)$, we can construct a different path from two points of γ within the normal neighborhood and this path will have longer proper time than the original γ. This method is known as cutting the corner. Therefore we have

$$0 = \left.\frac{dL}{dr}\right|_{r=0} = \frac{1}{c}\int_a^b g\left(w, \nabla_u u\right)\big|_{r=0} ds + \frac{1}{c} g(w,u)\big|_p. \tag{7.16}$$

If $u(a)$ is not orthogonal to Σ, then again we can find a normal neighborhood about $\gamma(a)$ and construct a different curve connecting points of Σ to γ which has greater proper time (we are using the fact that Σ has no boundary here). Thus $g(w,u)\big|_p = 0$ since w is tangent to Σ. Thus in order for $dL/dr\big|_{r=0} = 0$, we must have $\nabla_u u = 0$ along γ. Hence γ is a geodesic.

Conversely, if γ is a geodesic which is orthogonal to Σ, then $\nabla_u u = 0$ along $r = 0$ and $g\big(w(a), u(a)\big) = 0$. Therefore eq. (7.16) shows that $dL/dr\big|_{r=0} = 0$. □

Thus, when speaking of timelike curves with maximal length, we can restrict ourselves to variations of timelike geodesics which are orthogonal to spacelike hypersurfaces. The second variational formula for this scenario is the following proposition.

Proposition 7.19 *Let Σ be a spacelike hypersurface and $\gamma : [a,b] \to M$ a timelike geodesic which is orthogonal to Σ at $p = \gamma(a)$ normalized by $g(\gamma', \gamma') = -c^2$. Consider a timelike variation $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$ of γ such that $\alpha(a,r) \in \Sigma$ and $\alpha(b,r) = q =$*

$\gamma(b)$ for all $r \in (-\epsilon, \epsilon)$. Let w and u be the push forward of $\partial/\partial r$ and $\partial/\partial s$ under α and define the vector field v by $v^a = c^2 w^a + w^b u_b u^a$. Then

$$\left.\frac{d^2L}{dr^2}\right|_{r=0} = \frac{1}{c^5}\int_a^b g\big(v, \nabla_u(\nabla_u v) - c^2 R(u,w)u\big)\big|_{r=0} ds + \frac{1}{c^5}\sum_{i=1}^{k-1} g\big(v(s_i), \Delta_{s_i} v'\big) + \frac{1}{c^5} g(v, \nabla_u v)\big|_p + \frac{1}{c} g\big(L_{\gamma'}(w), w\big)\big|_p.$$

Proof. Notice that $v = 0$ at $q = \gamma(b)$ since $w = 0$ at q. Therefore

$$\sum_{i=0}^{k} g\big(v(s_i), \Delta_{s_i} v'\big) = \sum_{i=1}^{k-1} g\big(v(s_i), \Delta_{s_i} v'\big) + g(v, \nabla_u v)\big|_p.$$

So by proposition 7.9, it is only necessary to show

$$g\big(L_{\gamma'}(w), w\big)\big|_p = -g(u, \nabla_w w)\Big|_a^b = g(u, \nabla_w w)\big|_a = g(u, \nabla_w w)\big|_p.$$

But this is immediate from the definition of the second fundamental form operator. □

In the case that w is orthogonal to γ everywhere, then $v = c^2 w$ everywhere, so the second variational formula simplifies to the following.

Corollary 7.20 *Let Σ be a spacelike hypersurface and $\gamma : [a,b] \to M$ a timelike geodesic which is orthogonal to Σ at $p = \gamma(a)$ and normalized by $g(\gamma', \gamma') = -c^2$. If w is orthogonal to γ along γ, then*

$$\left.\frac{d^2L}{dr^2}\right|_{r=0} = \frac{1}{c}\int_a^b g\big(w, \nabla_u(\nabla_u w) - R(u,w)u\big)\big|_{r=0} ds + \frac{1}{c}\sum_{i=1}^{k-1} g\big(w(s_i), \Delta_{s_i} w'\big) + \frac{1}{c} g(w, \nabla_u w)\big|_p + \frac{1}{c} g\big(L_{\gamma'}(w), w\big)\big|_p.$$

This motivates the following definition of a Lorentzian index form for a spacelike hypersurface. Let $\gamma : [a,b] \to M$ be a timelike geodesic normalized by $g(\gamma', \gamma') = -c^2$

which is orthogonal to a spacelike hypersurface Σ at $\gamma(a)$. Assume that w is a piecewise smooth vector field along γ which is orthogonal to γ. If $w(a) \neq 0$ and $w(b) = 0$, then the ***index of*** w ***with respect to*** Σ is given by

$$I_\Sigma(w,w) = I(w,w) + \frac{1}{c} g\big(L_{\gamma'}(w), w\big)\big|_a$$

where

$$I(w,w) = \frac{1}{c}\int_a^b g\big(w, \nabla_{\gamma'}(\nabla_{\gamma'} w) - R(\gamma', w)\gamma'\big) ds + \frac{1}{c}\sum_{i=0}^{k-1} g\big(\Delta_{s_i} w', w(s_i)\big).$$

Hence $I(w,w)$ is just the usual Lorentzian index form. The partition $\{s_i\}$ of $[a,b]$ is chosen such that w is differentiable except at the s_i's.

Proposition 7.21 *Let $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$ be a timelike variation of the timelike geodesic γ, and assume the deviation vector $w\big|_{r=0}$ orthogonal to $\gamma(s)$ (i.e. $w\big|_{r=0} \in V^\perp(\gamma)$). Then*

$$\left.\frac{d^2 L}{dr^2}\right|_{r=0} = I_\Sigma(w,w).$$

Proof. This follows from the definition of $I_\Sigma(w)$ and corollary 7.20. □

So in order to show that focal points mark the end of timelike geodesics maximizing proper time, it suffices to find a variation α with deviation vector w such that $I_\Sigma(w,w) > 0$.

Theorem 7.22 *Let Σ be a spacelike hypersurface and $\gamma : [a,b] \to M$ a timelike geodesic which is orthogonal to Σ at $p = \gamma(a)$ and normalized by $g(\gamma', \gamma') = -c^2$. If $r = \gamma(s_1)$,*

with $s_1 \neq a, b$, is a focal point to Σ along γ, then there is a nontrivial *deviation vector field w which is orthogonal to γ, tangential to Σ, and satisfies$w(b) = 0$ and $I_\Sigma(w, w) > 0$. Consequently, there are timelike curves from Σ to $\gamma(b)$ which are longer than γ. Hence γ is not maximal.*

Proof. The proof is very similar to the proof of theorem 7.13. We seek a timelike variation α of γ which begins at Σ and ends at $\gamma(b)$ and satisfies $dL/dr\big|_{r=0} = 0$ and $d^2L/dr^2\big|_{r=0} > 0$. Proposition 7.18 shows $dL/dr\big|_{r=0}$ is satisfied for any timelike variation. By proposition 7.21 we want to find a timelike variation α with deviation vector w such that $w(a)$ is tangential to Σ, $w(b) = 0$, and $I_\Sigma(w, w) > 0$. But all we really need to do is construct a piecewise vector field w along γ such that $w(a) \in T_p\Sigma$, $w(b) = 0$, and $I_\Sigma(w, w) > 0$. This is because once we have this vector field w, we can construct the desired variation using proposition 7.6. The fact that $w(a)$ is tangential to Σ implies that the variation guaranteed by proposition 7.6 is indeed one that begins on Σ and ends at $q = \gamma(b)$. Then a timelike variation can be found by using proposition 7.5.

So our objective now is to construct a piecewise vector field w along γ such that $I_\Sigma(w, w) > 0$. By hypothesis there exists a nontrivial Jacobi field w_1 along γ such that w_1 is orthogonal to γ, vanishes at $r = \gamma(s_1)$, and satisfies $\nabla_{\gamma'} w_1 = -L_{\gamma'}(w_1)$ at p. We extend w_1 to $\widetilde{w}_1$ by

$$\widetilde{w}_1(s) = \begin{cases} w_1(s) & : \text{ for } a \leq s \leq s_0 \\ 0 & : \text{ for } s_0 < s \leq b. \end{cases}$$

Notice that since w_1 is nontrivial from a to b, we have

$$\Delta_{s_1}\widetilde{w}_1' = \lim_{s \to s_1^+} \nabla_u \widetilde{w}_1(s) - \lim_{s \to s_1^-} \nabla_u \widetilde{w}_1(s) = -\lim_{s \to s_1^-} \nabla_u w_1(s) \neq 0.$$

Since w_1 is orthogonal to γ and γ is a geodesic, it follows that $\lim_{s\to s_1^-} \nabla_u w_1(s)$ is orthogonal to γ' at s_1. Hence $\Delta_{s_1}\widetilde{w}_1'$ is nonzero and spacelike. Define a vector field $v \in T^\perp(\gamma)$ such that $v(a) = v(b) = 0$ and $g\big(v(s_1), \Delta_{s_1}\widetilde{w}_1'\big) = -1$. The existence of v follows from suitable smooth cut-off functions and the fact that $\Delta_{s_1}\widetilde{w}_1'$ is nonzero and spacelike. For $\epsilon > 0$, define

$$w_\epsilon = \epsilon^{-1}\widetilde{w}_1 - \epsilon v.$$

The index of w_ϵ with respect to Σ is given by

$$\begin{aligned}
I_\Sigma(w_\epsilon, w_\epsilon) &= I(w_\epsilon, w_\epsilon) + c^{-1} g\big(L_{\gamma'}(w_\epsilon), w_\epsilon\big)\big|_a \\
&= I(w_\epsilon, w_\epsilon) + c^{-1} g\big(L_{\gamma'}(\epsilon^{-1}\widetilde{w}_1 - \epsilon v), \epsilon^{-1}\widetilde{w}_1 - \epsilon v\big)\big|_a \\
&= I(w_\epsilon, w_\epsilon) + \epsilon^{-2} c^{-1} g\big(L_{\gamma'}(\widetilde{w}_1), \widetilde{w}_1\big)\big|_a \\
&= I(w_\epsilon, w_\epsilon) - \epsilon^{-2} c^{-1} g(\nabla_{\gamma'}\widetilde{w}_1, \widetilde{w}_1)\big|_a \\
&= \epsilon^{-2} I(\widetilde{w}_1, \widetilde{w}_1) + \epsilon^2 I(v, v) - 2I(\widetilde{w}_1, v) - \epsilon^{-2} c^{-1} g(\nabla_{\gamma'} w_1, w_1)\big|_a \\
&= \epsilon^2 I(v, v) - 2I(\widetilde{w}_1, v).
\end{aligned}$$

The third equality follows since $v = 0$ at p. The fourth equality follows since

$$g\big(L_{\gamma'}(\widetilde{w}_1), \widetilde{w}_1\big)\big|_a = g(\nabla_{\widetilde{w}_1}\widetilde{w}_1, \gamma')\big|_a = -g(\widetilde{w}_1, \nabla_{\widetilde{w}_1}\gamma')\big|_a = -g(\widetilde{w}_1, \nabla_{\gamma'}\widetilde{w}_1)\big|_a.$$

The fifth equality is just expanding out w_ϵ and recognizing that $\widetilde{w}_1 = w_1$ near a. The sixth equality uses the fact that since $\widetilde{w}_1$ is a Jacobi field and vanishes at r, we have

$$I(\widetilde{w}_1, \widetilde{w}_1) = c^{-1} g(\nabla_{\gamma'} w_1, w_1)\big|_a.$$

Now since $\widetilde{w}_1$ is a piecewise smooth Jacob field, we have

$$I(\widetilde{w}_1, v) = g\big(v(s_1), \Delta_{s_1}\widetilde{w_1}'\big) = -1$$

by construction. Thus

$$I_\Sigma(w_\epsilon, w_\epsilon) = \epsilon^2 I(v, v) + 2.$$

By taking ϵ small enough, we can find a w_ϵ such that $I_\Sigma(w_\epsilon, w_\epsilon) > 0$. □

7.5 Cosmological Singularities

We are now ready to prove singularities in spacetimes which model cosmology. The first theorem can be interpreted as showing that if the universe is globally hyperbolic and at an instant of time, the universe is expanding everywhere at a rate which is bounded from zero, then the universe must have had a beginning a finite amount of time in the past. These theorems were originally formulated by Stephen Hawking.

Theorem 7.23 *Let (M, g) be a globally hyperbolic spacetime with $R_{ab}u^a u^b \geq 0$ for all timelike u^a, which will be the case if the strong energy condition holds. Suppose there exists a smooth spacelike Cauchy Surface Σ and let L_n denotes its second fundamental form operator where n is the past directed timelike vector which is orthogonal to Σ and satisfies $g(n, n) = -c^2$. If $\theta = -\mathrm{tr}(L_n) \leq C < 0$ everywhere on Σ, then there is no past directed timellike curve from Σ which has proper time greater than $3/|C|$. Hence, all past directed timelike geodesics are incomplete.*

Proof. Suppose there exists a past directed timelike curve, λ, from Σ which has existed for proper time greater than $3/|C|$. Let $p \in I^-(\Sigma) \cap \lambda$ lie beyond proper time $3/|C|$. By theorem 6.46, there exists a maximal geodesic γ from p to q where $\{q\} = \lambda \cap \Sigma$,

which, of course, has proper time greater than $3/|C|$ (recall the length of γ is just c times its proper time). γ intersects Σ orthogonally from proposition 7.18 and by theorem 7.22, there is no point on γ which is a focal point to Σ. But proposition 7.17 implies that a focal point must exist on γ. This contradiction implies λ cannot exist. $\square$

It might seem more reasonable to conclude from theorem 7.23 that spacetime is not globally hyperbolic. However, the following theorem shows that even non globally hyperbolic spacetimes can still be singular as long as there is an edgeless, achronal, compact spacelike hypersurface. Unlike 7.23, we arrive at the much weaker conclusion that there is at least one incomplete timelike geodesic.

Theorem 7.24 *Let (M, g) be a spacetime with $R_{ab}u^a u^b \geq 0$ for all timelike u^a, which will be the case if the strong energy condition holds. Suppose there exists a smooth, edgeless, achronal, compact spacelike hypersurface Σ and let L_n denotes its second fundamental form operator where n is the past directed timelike vector which is orthogonal to Σ and satisfies $g(n, n) = -c^2$. If $\theta = -\mathrm{tr}(L_n) \leq C < 0$ everywhere on Σ, then there is at least one inextendible past directed timelike geodesic from Σ which has proper time no greater than $3/|C|$.*

Proof. Suppose that every past directed inextendible timelike geodesic from Σ has length greater than $3/|C|$. Since the spacetime $\big(\mathrm{int}\big[D(\Sigma)\big], g\big)$ satisfies the hypotheses of theorem 7.23, each past directed inextendible timelike geodesic must intersect $\partial D(\Sigma)$ and so they must all intersect $H(\Sigma)$ by proposition 6.16 and in particular they must all intersect $H^-(\Sigma)$, so $H^-(\Sigma) \neq \emptyset$.

The existence of an incomplete past directed timelike geodesic will follow from showing $H^-(\Sigma)$ is compact.

Thus it suffices to show $H^-(\Sigma)$ is compact. To do this, we will first show that for each $p \in H^-(S)$, there exists a timelike geodesic γ connecting Σ to p such that γ maximizes the proper time of all causal curves from Σ to p. To find γ, first notice that the proper time of any causal curve from Σ to $p \in H^-(S)$ is bounded above by $3/|C|$, so the supremum exists, τ^*, of the proper time of all causal curves from Σ to p exists. Let $\{\lambda_n\}$ be a sequence of timelike curves from Σ to p such that

$$\lim_{n\to\infty} L(\lambda_n) = c\tau^*$$

where L is the Lorentzian length functional. Choose $q_n \in \lambda_n$ such that $q_n \neq p$ but $\{q_n\}$ converges to p. Since $q_n \in I^+(p)$, we have $q_n \in \operatorname{int}\big[D^-(S)\big]$. By theorem 6.46 there is a geodesic γ_n from $r_n \in \lambda_n \cap \Sigma$ to q_n. Since Σ is compact, the sequence $\{r_n\}$ has an accumulation point r. Let $\{r_{n'}\}$ converge to r and let γ be the past directed geodesic starting at r and orthogonal to Σ which ends at $H^-(\Sigma)$. γ must end at p by continuity of the exponential map. By choosing $\{q_n\}$ to lie in $N \cap I^+(p)$ where N is a normal neighborhood of p, we can ensure that $L(\gamma) \geq L(\gamma_n)$ for each n. Also for each n, we have $L(\gamma_n) \geq L(\widetilde{\lambda}_n)$ where $\widetilde{\lambda}_n$ is the restriction of λ_n from r_n to q_n. These observations imply

$$L(\gamma) \geq \lim_{n'\to\infty} L(\gamma_{n'}) \geq \lim_{n'\to\infty} L(\widetilde{\lambda}_{n'}) = c\tau^*$$

which implies $L(\gamma) = c\tau^*$ since τ^* is the supremum of proper time over all timelike curves from Σ to p. Thus, we have shown that there exists a timelike geodesic γ connecting Σ to p such that γ maximizes the proper time of all causal curves from Σ to p.

Now we finally show $H^-(\Sigma)$ is compact. Let $\{p_n\}$ be a sequence in $H^-(\Sigma)$. Let γ_n

be the geodesic orthogonal to Σ which maximizes the proper time of all causal curves from Σ to p. Let r_n be the intersection of γ_n with Σ. Then $\{r_n\}$ is a sequence of points in Σ so there exists an accumulation point $r \in \Sigma$. Let γ be the past directed geodesic starting at r and orthogonal to Σ which intersects $H^-(\Sigma)$ at some point p. Using the continuity of the exponential map, we find that p is an accumulation point of $\{p_n\}$.

If we assume (M, g) is strongly causal, then we can immediately find a contradiction: $H^-(\Sigma)$ contains a future inextendible null geodesic by theorem 6.15 since $\text{edge}(\Sigma) = \emptyset$. This contradicts proposition 6.18 since $H^-(\Sigma)$ is compact.

However, we do not need to assume (M, g) is strongly causal to obtain a contradiction. For each $p \in H^-(\Sigma)$ let γ_p denote a timelike geodesic connecting Σ to p which maximizes the proper time of all causal curves from Σ to p. Define the function $f : H^-(\Sigma) \to \mathbb{R}$ by $f(p) = L(\gamma_p) = \sup_{q\in\Sigma} d(p, q)$ where d is the Lorentzian distance function introduced in section 6.7. By replicating proposition 6.41, we can show that f is lower semi-continuous. Since f is lower semi-continuous on the compact set $H^-(\Sigma)$, it must obtain its minimum at some point $p_0 \in H^-(\Sigma)$. Construct a future-directed null geodesic $\lambda : [a, b] \to M$ such that $\lambda\big([a, b]\big) \subset H^-(\Sigma)$. The existence of such a λ is guaranteed by theorem 6.15. Fix $s > a$. Our desired contradiction will follow from showing $f\big(\alpha(s)\big) < f(p_0)$. Let σ denote the future directed path from p_0 to Σ by first following α until $\alpha(s)$ and then following $\gamma_{\alpha(s)}$. σ is the union of a null geodesic and timelike geodesic, so it's piecewise differentiable. While keeping the endpoints of σ fixed, we deform σ around the non-differentiable point $\alpha(s)$ into the causal curve $\widetilde{\sigma}$ which has

length greater than σ. Then

$$f\big(\alpha(s)\big) = L(\gamma_{\alpha(s)}) = L(\sigma) < L(\widetilde{\sigma}) \leq L(\gamma_{p_0}) = f(p_0),$$

which contradicts p_0 being the minimum point of f. □

7.6 An Almost Realistic Singularity-Free Cosmological Model

In this final section, we briefly present a singularity-free model which can almost represent our universe. The importance of this model shows that it still may be possible to construct physically, realistic cosmological models that can describe our universe and yet do not possess singularities. Theorems 7.23 and 7.24 leave little room to do this. By ***singularity-free***, we mean null and timelike geodesically complete and by ***realistic cosmological model***, we mean a spacetime that can adequately describe the observed homogeneity, isotropy, and expansion of our universe. Such a singularity-free model has not been found. However, the model we present here is singular-free and can describe the expansion of the universe and satisfies all energy condition. Unfortunately, it does not describe the homogeneity (and therefore the isotropy) of the universe. Nonetheless, this example shows that there might still be wiggle room within theorems 7.23 and 7.24 to find realistic cosmological models without any singularities. The example is do to José Senovilla (see [8]).

The spacetime is globally hyperbolic with topology $\mathbb{R}^4$. In cylindrical coordinates

$\{t, r, \phi, z\}$, the metric can be written as

$$\begin{aligned} g = &\cosh^4(ct)\cosh^2(3r)(-c^2dt^2 + dr^2) \\ &+ \frac{1}{9^2}\cosh^4(ct)\cosh^{-2/3}(3r)\sinh^2(3r)d\phi^2 \\ &+ \cosh^{-2}(ct)\cosh^{-2/3}(3r)dz^2. \end{aligned}$$

We see that $\frac{\partial}{\partial\phi}$ and $\frac{\partial}{\partial z}$ are Killing vectors, so the spacetime possesses cylindrical symmetry. This metric is a solution to Einstein's field equations with a stress-energy tensor of a perfect fluid:

$$T_{ab} = (\rho + c^{-2}P)u_a u_b + P g_{ab}$$

where

$$\rho = \frac{15c^4}{8\pi G}a^2\cosh^{-4}(ct)\cosh^{-4}(3\rho) \qquad \text{and } P = \frac{c^2\rho}{3}.$$

This relationship between P and ρ is expected to hold for the early history of the universe when the energy was mostly dominated by radiation. This is the main reason for the random appearance of 3's in the metric. Examining ρ and P, we see that this spacetime obeys the weak, strong, and dominant energy conditions. The four-velocity of the fluid is given by

$$u = -\cosh^2(ct)\cosh(3r)\frac{\partial}{\partial t}.$$

The fluid is orthogonal to the spacelike hypersurfaces defined by constant t, hence the rotation tensor of the fluid congruence satisfies $\omega = 0$ by lemma 7.1. More importantly, the expansion satisfies

$$\theta = \nabla_a u^a = \frac{\sinh(ct)}{\cosh^3(ct)\cosh(3r)}.$$

Thus the universe is contracting for half its history ($t < 0$) and expanding for its other half ($t > 0$). An analysis of the geodesic equation shows that the spacetime is geodesically complete (hence singularity-free) and the spacelike hypersuraces of constant t are Cauchy surfaces. Thus this spacetime is globally hyperbolic.

Why doesn't theorem 7.23 apply to this spacetime? We see that on each spacelike hypersurface of constant t, $\theta \to 0$ as $r \to \infty$. Thus we don't satisfy the condition $\theta \leq C < 0$ in theorem 7.23. Therefore the hypothesis of being able to bound the expansion away from 0 is necessary for theorems 7.23 and 7.24. Thus this is the condition one must break in order to find realistic cosmological models.

8 References

[1] Beem, John K., Ehrlich, Paul E., and Easly, Kevin L., *Global Lorentzian Geometry*, second edition. Marcel Dekker Inc., New York, 1996.

[2] do Carmo, Manfredo Perdigão. *Riemannian Geometry.* Brikhäuser, Boston, 1992.

[3] Hawking, Stephen and Ellis, George *The Large Scale Structure of Space-Time.* Cambridge University Press, Cambridge, 1973.

[4] Lee, John M. *Introduction to Smooth Manifolds.* Springer, New York, 2013.

[5] Lee, John M. *Riemannian Manifolds: An Introduction to Curvature.* Springer, New York, 1997.

[6] Penrose, Roger. *Techniques of Differential Topology in Relativity.* Philadelphia, Siam, 1972.

[7] Senovilla, José M. "Singularity Theorems and Their Consequences." *General Relativity and Gravitation* **30** 701-848, 1998.

[8] Wald, Robert M. *General Relativity.* University of Chicago Press, Chicago, 1984.